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

By ARLEN GUARIN, CHRISTIAN POSSO, ESTEFANIA SARAVIA, JORGE TAMAYO [*]

February 21, 2021

Abstract

We estimate the effect of physicians on health outcomes by exploiting a Colombian government program that randomly assigns newly graduated physicians to hospitals across the country. Using administrative data from the program, vital statistics records, and individual records from the mandatory field-specific college graduation exams, we show that more-skilled physicians improve health at birth of infants whose mothers received care in those hospitals during their pregnancy. We show that the mechanisms underlying the results are the time physicians spend with the patient and their success in targeting care toward the most vulnerable patients.

Keywords: Physicians' skills, health birth outcomes, experimental evidence
JEL Codes: H51, I14, I15, I18

[*] Guarin: University of California, Berkeley (email: aguariga@berkeley.edu); Posso: Banco de la Republica de Colombia (email: cpossosu@banrep.gov.co); Saravia: Harvard University, Harvard Business School (email: esaravg@gmail.com); Tamayo: Harvard University, Harvard Business School (email: jtamayo@hbs.edu). The opinions expressed herein belong to the authors and do not necessarily reflect the views of Banco de la Republica or its Board of Directors. We thank Manuela Cardona, along with Santiago Velasquez, Silvia Granados, Gabriel Suarez, Nicolas Mancera and Brayan Pineda for excellent research assistance. We thank Maria Aristizabal, Francesco Bogliacino, Leonardo Bonilla, David Card, Kaveh Danesh, Margarita Gafaro, Robert Gonzalez, Hilary Hoynes, Rembrand Koning, Juliana Londono, Edward Miguel, Paul Rodriguez, Emmanuel Saez, Molly Schnell, Christopher Walters, Danny Yagan and seminar participants at Banco de la Republica, ESSEN health conference 2020, EEA congress 2020, Universidad Eafit, Universidad del Rosario, Development and Labor Lunch (at UC Berkeley) for insightful comments. We also thank the Colombian Ministry of Health, the *Administrative Department of Statistics* - DANE and the *Instituto Colombiano para la Evaluación de la Educación* - ICFES for providing access to the data and insighful discussions.

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 be predictive of future abilities and health trajectories that cannot be explained by genetics (Almond et al., 2005; Currie, 2011; Currie and Almond, 2011). In trying to understand the causes of such differences in birth outcomes, most of the literature has focused on parents' decisions during pregnancy, families' socioeconomic conditions (Currie, 2011), environmental factors (Currie and Schwandt, 2016b), and access to the health system in the extensive margins (Currie and Gruber, 1996; Finkelstein et al., 2012) and intensive margins (Almond et al., 2010). Yet an unresolved important question is whether more-skilled healthcare professionals can improve health outcomes at birth.

In this paper, we break a new ground by providing causal evidence on the role that skilled physicians play in newborns' health at birth. Studying this matter is important because physicians are arguably the health professionals who make the greatest contribution to patient health (Das and Hammer, 2005) and can affect investments in utero that determine infants' health at birth. Moreover, 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).

The lack of causal evidence regarding physicians' effect on birth outcomes is not surprising, since answering this question poses a substantial empirical challenge. It requires to account for the selection bias associated with the match between physicians and hospitals or patients (Doyle et al., 2010).[1] We overcome this challenge by exploiting a Colombian government program that randomly assigned 2,365 physicians to 592 small hospitals. We estimate the impact on the 104,358 children whose mothers received care in those hospitals during their pregnancy, using administrative data from the program, vital statistics records, and individual records from mandatory field-specific college graduation exams.

We leverage data available on teams of newly graduated physicians in Colombia. Colombia requires medical school graduates to work for the first year of their career in the national Mandatory Social Service (SSO), which randomly assigns them to hospitals across the country. We combine several rich granular administrative records and collect data on the reports

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

published by Colombia's Ministry of Health after the SSO lotteries.[2] There are few measurements of the skills of recent medical graduates; we use individual records from the mandatory field-specific college graduation exams to measure physician skills. Finally, we link the hospitals to which doctors were randomly assigned to the national Vital Statistics Records (VSR), from which we obtain birth outcomes and maternal sociodemographic characteristics.

Our random assignment setting has many advantages. For example, hospitals' characteristics do not covariate with physicians' skills, and new physicians arriving at a specific hospital have access to the same facilities as well as the same administrative and health staff. Also, by comparing across hospitals, we can estimate the causal effect physicians have on patients' health outcomes.

We find that an increase of one standard deviation in the college graduation exam scores of the team of physicians assigned to a hospital decreases the probability of low birth weight by 7.4%. These effects are consistent across alternative health measures at birth: we also find a negative impact of 10% on prematurity and a 13.7% decrease in the probability of low Apgar scores.[3]

To shed light on the potential channels through which physicians can impact child outcomes, we first analyze several heterogeneous effects across different mothers' characteristics. Although the effects are slightly more pronounced among first-time mothers,[4] mothers with low education, and married women, the differences between groups are not statistically different. We then estimate effects separately for male and female newborns. It has been well established that in utero, males are more vulnerable to health shocks than females (Eriksson et al., 2010; Kraemer, 2000; Naeye et al., 1971; Pongou et al., 2017). We investigate whether skilled physicians help to mitigate adverse shocks in utero. Although the reduction in low birth weight was particularly pronounced among male newborns, we do not find any statistical difference between male and female newborns.

Further, using health measures at the hospital level for the years before the SSO cohorts we consider in this paper, we study heterogeneous effects between hospitals with high and low incidence of poor newborn health during the three years before the cohorts we analyze. The effects on low birth weight are driven by hospitals with a high incidence of poor newborn health, which we define as the hospitals in the top quartile of the low birth weight baseline

[2] We focus on the lotteries that took place between 2013 and the third quarter of 2014.

[3] 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; for example, Almond et al. (2010) and Lin (2009).

[4] Similar to a large fraction of the literature, main estimates throughout the paper are based on first-time mothers.

incidence distribution. We find that an increase of one standard deviation in college graduation exam scores decreases low birth weight by 11.7% in hospitals with a high incidence of poor health at the baseline. Similar results are found in prematurity and Apgar score. Thus, more-skilled physicians improve low birth weight, prematurity, and Apgar score, but these effects are driven by hospitals with a high (pre-program) incidence of poor newborn health.

We next explore two key mechanisms through which physicians may improve health at birth. First, we study the time physicians spend with the mother during pregnancy. We split the data between teams of physicians for whom the mother's gestation period falls completely within the 12-month period that the physician spent in the hospital and the complement (i.e., mothers who had only partial exposure). We find much stronger effects in cases when the mother has more exposure time to the physicians.

Second, we explore the role of the number of prenatal consultations. According to WHO (2016) and the Colombian government (Gomez et al., 2013), better and more frequent prenatal care during pregnancy can improve the health of both the mother and her newborn.[5] We follow the standard recommended by WHO (2016) in 2013 and define "adequate prenatal care" as having at least four visits to the doctor during pregnancy.[6] We find that more-skilled doctors, on average, do not schedule more prenatal checkups.[7] We then test whether the more-skilled physicians target prenatal consultations toward the most vulnerable mothers, measured as those with predicted likelihood of giving birth to a baby with low birth weight. We use several machine learning techniques to generate two groups of predictions about mothers' low birth weight probability.[8] We use a set of mother-hospital characteristics available for physicians at the time of prenatal care. The results show how, regardless of the method we use to predict low birth weight, more-skilled doctors do not increase the suggested number of antenatal consultations with mothers with a low predicted probability of low birth weight. Still, they seem to target those prenatal checkups toward

[5] 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 in utero infant growth (Amarante et al., 2016; Kramer, 1987). Further, in Colombia, the Ministry of Health requires that prenatal checkups be carried out by physicians (Gomez et al., 2013); thus, physicians are responsible for prenatal care, and they are the professionals who attend 98% of deliveries.

[6] In our sample, 87% of mothers have at least four visits to the doctor.

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

[8] We apply four algorithms: random forest, XGBoost (Chen and Guestrin, 2016), neural network (Hoffman et al., 2018), and logistic regression models.

more vulnerable mothers, measured as mothers with a higher predicted probability of low birth weight. Finally, we show that the effects on birth outcomes are particularly pronounced among mothers with an ex ante high predicted probability of low birth weight. Altogether, these results are consistent with physicians being time constrained and unable to increase the average amount of time spent in 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. First, we assign a placebo treatment to infants born before the arrival of the physicians in our sample. We run placebo tests similar to our main specification using data for the three previous years (2010-2012) for which the doctors working at hospitals were randomly assigned (2013-2015). We find that the treatment generates precisely estimated zeros. Second, we run similar estimation procedures using the municipality and hospital characteristics before the physicians' arrival as a dependent variable and find no significant relationship. Altogether, we read these results as evidence of the randomness of the assignment of physicians to hospitals.

Our identification strategy and the availability of granular administrative records allow us to contribute to several strands of the literature. First, we contribute to the literature on physicians' effects on health outcomes. This literature documents the relationships between health outcomes and healthcare costs (Alsan et al., 2019; Clemens and Gottlieb, 2014; Molitor, 2018), quality of physicians' academic institutions (Doyle et al., 2010), physicians' performance on qualifying examinations (Carrera et al., 2018; Tamblyn et al., 2002; Wenghofer et al., 2009), physicians' competence (Das et al., 2016)[9], physicians' ability to facilitate adherence to prescription medications (Iizuka, 2012; Simeonova et al., 2020), physicians' fees and payment for performance (Basinga et al., 2011; Ho and Pakes, 2014a,b), general practitioners and specialists (Baicker and Chandra, 2004), and physicians' communication (Curtis et al., 2013). To our knowledge, our paper is the first to document experimental evidence of the impact that more-skilled physicians have on health at birth outcomes.

Our research is also related to the literature that studies the effects of healthcare access on health outcomes (Almond et al., 2010; Finkelstein et al., 2012).[10] In particular, our paper relates to Currie and Gruber (1996), who show that access to health insurance for pregnant women lowered the incidence of low birth weight.

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

[10] See Aron-Dine et al. (2015), Bardach et al. (2013),Michalopoulos et al. (2012), Anderson et al. (2012), Anderson et al. (2014) for studies related with the effects of healthcare access on population health.

Our study adds to the literature on overuse and inefficient resource allocation by physicians and hospitals (Abaluck et al., 2016; Chandra and Staiger, 2020; Currie and MacLeod, 2017). In particular, Abaluck et al. (2016) show that physicians do not target testing to the highest-risk patients, since observable risk factors receive little attention in physicians' testing decisions. In this paper, we take advantage of recent advances in machine learning techniques to show that more-skilled physicians target prenatal consultations toward mothers with the highest risk of low birth weight.

We add to the large body of research that has studied the origins of inequality at birth (Black et al., 2007; Chetty et al., 2011; Currie, 2011) and how heterogeneity of endowments at birth affects future outcomes such as earnings, education, and health (Currie, 2009; Oreopoulos et al., 2008; Persson and Rossin-Slater, 2018). We provide new evidence by showing that children born under the care of less knowledgeable physicians are indeed more likely to exhibit worse health at birth.

Finally, our paper is related to the literature on teacher value added, where the effect on students of a high-quality (effective) teacher has proved to be significant (Araujo et al., 2016; Chetty et al., 2011; Rivkin et al., 2005; Rockoff, 2004). While this literature 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 one standard deviation increase in physicians' quality decreases the probability of low birth weight by 7.4 percent. Our findings suggest that, similar to teachers, good doctors have the potential to effect great social value through better child outcomes at birth.

The rest of the paper is organized as follows: In Section 2, we describe the Colombian health system and the SSO program, the setting that we exploit to identify parameters of interest. Section 3 describes the rich administrative data we derive from doctors' college exit exams and patients' outcomes at birth. In Section 4, we introduce our estimation strategy, while in Section 5, we show evidence on the randomness of physicians' assignment to hospitals and present our main estimated effects. Section 6 discusses potential mechanisms, and Section 7 concludes.

2 Institutional Background and the Experimental Setting

2.1 Institutional Background

According to Colombia's political Constitution, 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 of 1993). Law 100 of 1993 introduced two types of health insurance: subsidized and contributive. The contributive regime is made up of formal employees (and their families) who contribute a fixed share of their employment income to the system. The subsidized regime is made up of poor household members who do not have formal employment.[11] By 2011, the access to health-care was close to universal, and even in the poorest population, the coverage was 87%, while in rural areas it was about 88% (Páez et al., 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 in Colombia is 87.7%, while the percentages of newborn 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 database[12], 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.

An important characteristic of the Colombian health system is that prenatal examinations must be carried out by physicians. According to the practical guide for the prevention, early detection and treatment of 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[13] and, in fact, calculations from the VSR show that physicians are responsible for all prenatal check-ups and 98% of deliveries are attended by physicians.

To become a physician in Colombia, one must study an undergraduate program in medicine. Similar to the college programs in nursing, bacteriology, and dentistry, medicine is considered a health program. Students accepted into health programs earn a BA after five to

[11] 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.

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

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

six years of education. According to Colombian law, all professionals graduating from health programs are social servants, and right after graduation, they must provide professional services in urban and rural areas lacking 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, but it was only adopted by 2010, when its implementation was legislated by Resolution 1058/2010. Besides the objective of improving access and quality of health services in depressed urban and rural populations or those with difficult access to health services and stimulating an adequate geographical distribution of human talent in health, the SSO also targets the promotion of spaces for the personal and professional development of those beginning their careers in the health sector.[14]

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 which hospitals they were assigned became one of the program's most critical decisions. In regard to this decision, the Law 1164/2007 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 way of running the random assignments was introduced. The first years of implementation of the new SSO program proved that the directions given by Resolution 1058/2010 were not strong enough to guarantee a transparent and organized assignment of physicians. Resolution 4503/2012 was introduced to give clearer and more organized guidance about how the random draws should be conducted. Resolution 566/2012 mandated that starting in January 2013 there would be four yearly waves of SSO draws[15] where professionals who applied to a specific state would be randomly assigned to the available positions in that state. To avoid strategic application behavior and to take advantage of the fact that the number of newly graduated physicians was around two times 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 still lower than two times the number of available places. The aforementioned feature of the process about the number of available places guaranteed an excess of demand for spots in each state

[14] See resolution 1058/2010.

[15] Taking place in January, April, July, and October in each of the 32 states.

and cohort.

After the application process is closed, 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 through the use of ballots. If a health professional gets a white ballot, then they are exempt of the social service and will receive a certificate that allows them to work in Colombia as a professional. Otherwise, the professional gets a red ballot with the code of the specific hospital where they will provide their services as professionals. If there are fewer professionals than spots available, all professionals registered are assigned to a hospital. Still, the specific hospital is assigned through the lotteries. Finally, the civil servant of the secretariat of health prepared a report stating the winners and their assigned hospitals, as well as the professionals who are exempt from the SSO program.

The social service at the assigned hospitals begins around one or two months after the draw and lasts for 12 months. If a health professional refuses to work in the place they were assigned to or unilaterally quits before the official end of the service, they are given a six-month sanction where they cannot work as health professionals. After that period, they have to 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.[16] The system of assigning professionals to hospitals randomly lasted for seven draws.[17] 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 randomly selected and the communities that get assigned doctors with various skills. The latter are the focus of this paper; the implications for the professionals are studied in Guarin et al. (2021). We use the exogenous rule of assignment to compare the birth outcomes of patients in hospitals who were assigned professionals with different skills but are otherwise comparable. In this paper, we focus on birth outcomes, given the relevance of these variables for future human development, and on medium- and long-term

[16] We cannot confirm whether physicians did actually work for the hospitals they were assigned to, but using information from payments to the social security system, we observe that 80% of the winners got a job as physicians after the draw. This gives us a measure of the level of compliance of the program.

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

inequalities.

Despite the SSO being mandatory for health graduates of different fields[18], 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 clearer, 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) and in particular to birth outcomes.

3 Data

We use data from four main administrative records. The primary dataset comes from the reports written after each state-level draw and published by the Ministry of Health for the draws 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 that the physician applied to, whether the student "won" the lottery or not, and notably the hospital to which each was randomly assigned and the proposed start date. For our period of analysis, 45 % of the hospitals in the program show up in only one draw, while 29% of the hospitals appear in two draws and 26% of the hospitals 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, 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 post medical training exams (SABER PRO). From the SABER PRO, we get physicians' individual performance in four different fields, including reading (comprehension), quantitative (reasoning), public health, and health management, plus some detailed sociodemographic information about each professional.

In our estimations, we use the scores in the four fields as proxies of the physician's skills 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

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

knowledge in mathematics, statistics, and data analysis. The specific medical competencies are included in the public health and health management modules. The health management module evaluates competencies related to understanding administrative processes, aspects including planning, organization, management, and control of health services. In particular, it evaluates the understanding of administrative processes for the development of health activities, recognition of services provided to the patient in the legal framework, and application of patient safety and ethical standards in the provision of health services. The public health score measures basic concepts regarding prioritized treatment plans for individual patient conditions. Essentially, the health module tests the physician's knowledge of components and processes of primary health care. In addition, this test is designed to recognize the treatments related to health conditions and to apply them in the selection of intervention actions for potentially basic medical conditions.

Our main exercises use the average score and the first principal component of the four fields; nonetheless, the general results do not change when we use each score individually or different ways to summarize the variation in the different fields. Since 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 physicians' general and medical skills at the time they start their SSO service and professional career. [19]

In Colombia, as many other developing countries, there is high heterogeneity in the 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 et al., 2011). Figure 1 shows high heterogeneity on the average score of the SABER PRO test between and within programs (and universities) for the physicians in our sample.[20] The Figure shows the mean score for each university/program and an interval of one standard deviation to each side of the average. 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 since it allows us to compare the outcomes of patients who were randomly exposed to physicians with very different knowledge bases and skills.

Using the scores and demographic characteristics from the SABER PRO, Guarin et al. (2021) show that the SSO lotteries in our sample are well balanced between winners and losers. Table A.1 and Figure A.1 replicate the balancing tests in Guarin et al. (2021). Table A.1 shows individual regression between the lotteries and physicians' characteristics. In

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

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

addition, Figure A.1 uses machine learning techniques and a classification permutation test to provide evidence of equality of multivariate distributions between treatment and control groups (Gagnon-Bartsch et al., 2019).[21]

The third administrative dataset comes from the 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 hospitals within Colombia's 1,120 municipalities from 1998 to 2018. Using hospitals' identification codes, we are able to link physicians and the birth records of the hospitals that they were assigned to. Using the birth date and number of gestation weeks from VSR, we are able to identify children born between 2013 and 2015 who were exposed to each team of physicians. We also use the VSR data from 2010 to 2012 to create mother and hospital-level controls to provide evidence of the covariate balance at the hospital level and to run falsification tests (Administrative Department of Statistics, 2018).

Finally, 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 control variables at the municipality level. We also collected additional data at the hospital level from the Colombian Ministry of Health.

3.1 Main sample

Our primary data source are the draws implemented in January, April, July, and October 2013 and January, April, and July 2014. Since the objective of the program is to provide professional services in urban and rural areas with difficult access to health services, and given that 77.3% of the available positions in these draws were located in small cities outside of the main 23 Colombian metropolitan areas, in this exercise we exclude municipalities in metropolitan areas, where we expect assigned physicians to play a less pivotal role.[22] The municipalities included in our sample cover around 58% of the Colombian population. We further constrain our sample to hospitals with at least one physician assigned in the seven draws and at least one birth certificate filed from 2013 through 2016. Finally, our main estimates focus on first-time mothers, although we also show estimates for non-first-time mothers.

For each newborn, we observe the complete birth certificate, which includes information

[21] We also perform a simple reverse regression to show that the set of baseline covariates do not explain the treatment variable. We found no evidence in this matter (test F(19,160) = 0.88, p-value=0.6128).

[22] In the appendix, Table A.2, we show that our main results hold when we include all municipalities.

Figure 1: Heterogeneity in Saber Pro scores in medicine programs

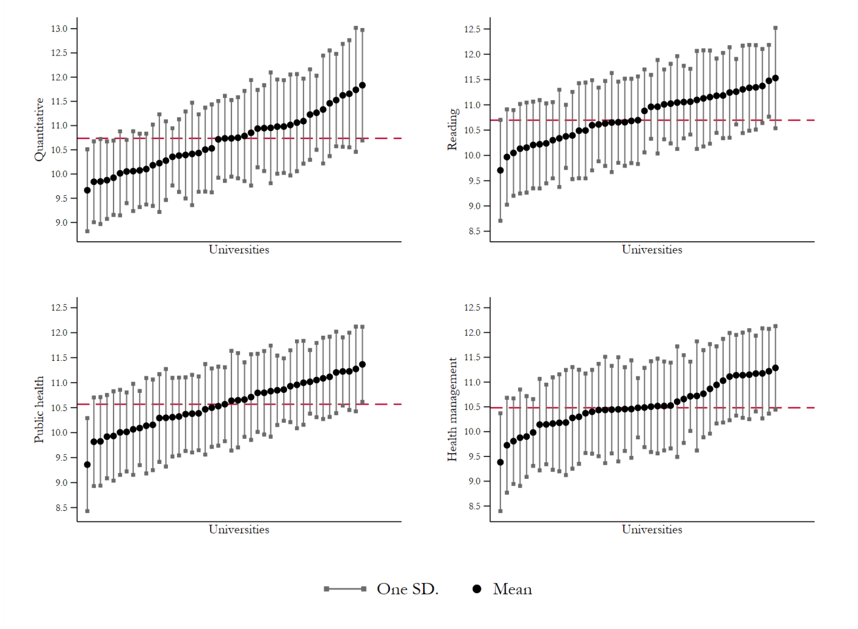


Figure 1 reports the reading, quantitative, health management, and public health test scores for the universities that the physicians in our sample attended. Data accounts for 44 different universities. The figure shows the mean score for each university/program and an interval of one standard deviation 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.

on low birth weight, Apgar score, weeks of gestation, and demographic information for mother and newborn. For each physician, we observe the four scores in reading, quantitative, public health, and health management, plus some socio-demographic information. Our final sample contains 104,358 combinations of newborns and physicians.[23]

Table 1 provides the basic descriptive statistics for the main variables used from the VSR. It also shows how our sample changes as we add the restrictions used in our main estimations. The first column shows the mean for newborns in hospitals where at least one SSO physician was assigned (SSO sample); column 2 shows the same statistics when we constrain the sample to first-time mothers only; and column 3 shows the mean when we further constrain the sample to the municipalities outside of the main metropolitan areas. The last column corresponds to our final main sample. In our main sample, 4.9% of births

[23] Several physicians were assigned to metropolitan areas and others to hospitals that did not filed a birth certificate from 2013 through 2016.

were low birth weight, 4.2% were early-term infants, and 4.6% of births had an Apgar score below 7. Finally, teenage pregnancy is 54.7% of total births in the main sample, compared with 28.4% of the total births in the SSO sample.

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

Covariate	SSO sample		SSO sample constrained to first-time mother		SSO sample constrained to first-time mother and non-MA	
	Mean	SD	Mean	SD	Mean	SD
Low birth weight	0.0601	0.2377	0.0675	0.2509	0.0494	0.2168
Very low birth weight	0.0061	0.0780	0.0062	0.0782	0.0035	0.0591
Prematurity	0.0623	0.2417	0.0625	0.2420	0.0415	0.1994
Apgar 1 min <7	0.0378	0.1908	0.0446	0.2064	0.0463	0.2101
Prenatal visits $\geq$ 4	0.8202	0.8202	0.8202	0.8202	0.8202	0.8202
Female newborn	0.4877	0.4998	0.4866	0.4998	0.4870	0.4998
Teenage mother	0.2840	0.4509	0.5300	0.4991	0.5469	0.4978
Number of observations	372,609		155,745		104,358	

Notes: This table presents the mean and standard deviation (SD) of the main birth statistics of the newborns affected by the SSO program. The data comes from the 2013-2016 DANE VSR, which collects and provides information that reveals the changes in mortality and fertility. 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 1 is the proportion of newborns whose Apgar 1 score is lower than 7; Female newborn is the proportion of female newborns. Prenatal visits $\geq$ 4 is the proportion of mothers who had at least four visits; and teenage mother is the proportion of mothers aged 19 years old or less.

4 Empirical Strategy

Our empirical setting focuses on a health production function that relates health outcomes at birth to physicians' skills. In our setting, multiple teams were randomly assigned to a large number of patients who are associated with a specific hospital. 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 a more-skilled physician on a newborn's health outcome (e.g., low birth weight, prematurity), using the following linear specification:

$$Y_{h,t,i} = \alpha + \gamma_t + \beta Z_{h,t} + X'_{h,t}\Phi + W'_{h,t,i}\Theta + \epsilon_{h,t,i} \tag{1}$$

where $Y_{h,t,i}$ is the outcome of child i born in hospital h and exposed to a physician team t. $Z_{h,t}$ is a score that measures the overall skills of the physician team t that was randomly assigned to serve in hospital h and whose service period intersects with child i's gestation period.[24] X_h is a vector of ex ante hospital and team characteristics. W_i is a vector with sociodemographic information of mother-child i, and γ_t are draw-by-state fixed effects.[25] Because physicians in each draw and state are assigned to hospitals, controlling for draw-by-state fixed effects (γ_t) is crucial to our identification strategy; otherwise, variation in physician quality could reflect other regional differences in the assignment of physicians to hospitals. Finally, standard errors are clustered by team and child.

The coefficient of interest is β. Under the assumption that teams of doctors within each draw state were randomly assigned to hospitals, the β estimated by OLS cleanly identifies the effect of a more skilled team of physicians on children potentially exposed to their service in the assigned hospital h. To make the interpretation of the estimated coefficient β straightforward, we divide $Z_{h,t,i}$ by its standard deviation and relate the effect to the average of the outcome. Therefore, the final result is interpreted as the percentage change in the outcome variable associated with one standard deviation increase in the skill measure. We also estimate heterogeneous effects using the demographic characteristics of the newborns, their mothers, and the hospitals in which they were born.

To evaluate the internal validity of our identification strategy, we implement the following falsification tests. We assign "placebo treatment" to the newborns who show up in the VSR of the three years before the program (2010, 2011, and 2012) instead of years 2013, 2014, and 2015 used in our main estimation sample. We use the same draw date, proposed start date, and hospital to which each of the physicians was randomly assigned but three years before the actual date. We then run equation (1) under the same conditions used for the main sample.

Like most literature in economics, we focus on low birth weight as our principal measure of health at birth (Currie, 2011). We use prematurity and the Apgar score as related measures of health at birth and to provide robustness for our main estimations.[26] Prematurity is

[24] We explore two different measures as proxies of the physicians' skills: the average score and the first principal component of the four tests available. The results are robust to this decision.

[25] Since we use data for three years of the infants' vital statistics, we also include year fixed effects to control for changes over time. Sociodemographic characteristics of the newborns and their mothers include infants' gender and mothers' age, education, access to subsidized health, and marital status.

[26] 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. Complications include immunological, respiratory, central nervous system, gastrointestinal, hearing, and vision problems as well as cognitive, motor, social-emotional,

defined as being born before the 37th gestational week. For Apgar, we use an indicator of whether the newborn had a score below 7 in Apgar 1, as the threshold of 7 is commonly used in the literature (Ehrenstein, 2009). Almond et al. (2005) argue 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). Similarly, Ma and Finch (2010) recommend always including the Apgar score since it appears to be the strongest predictor of neonatal mortality, regardless of birth weight.

We focus on the average score for most of our analysis. Nonetheless, we provide robustness results using the first principal component and each score individually. 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.[27] Finally, for the entirely of the analysis, we focus on our main sample. In the Appendix, as a robustness check, we present the results for other samples.

5 Results

This section describes the causal effects of physicians' skills on birth outcomes. We first test whether hospitals' birth outcomes and additional covariates measured in years 2010, 2011 and 2012 from VSR (randomly assigned) arrival are correlated with their score. Our results show that there is no correlation between different health outcomes and our proxy for physicians' skills. Second, we find that physicians' skills have a negative and significant effect on low birth weight, prematurity, and Apgar. Third, we provide robustness checks to our main results by using a standardized principal component as a proxy for physicians' skills and excluding the controls, and using different functional forms. Fourth, we implement a placebo test.

behavioral, and long-term growth problems (Butler et al., 2007; Currie and Walker, 2011; Taylor et al., 2001; Veddovi et al., 2001). Callaghan et al. (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. Apgar has also been used in the literature as a measure of newborn health status; for example, 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 et al., 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 et al., 2002). Among school-age children, low Apgar scores are also associated with minor language, motor, speech, and developmental impairments (Razaz et al., 2016).

[27] Our results hold when we use an unweighted average of the scores.

Fifth, we estimate heterogeneous effects on mothers' and hospitals' characteristics. Finally, we explore prenatal consultations as a mechanism to improve the quality of care and health outcomes.

5.1 Characteristics of the hospitals and physicians' skills

To test whether the main health at birth outcomes and additional covariates, measured before the program, are correlated with the quality of the physicians assigned to each hospital, we regress each hospital's characteristics three years before the SSO program on physicians' average college examination scores. We include date and state (where the draws took place) fixed effects and cluster the standard errors by hospital. Table 2 shows the coefficients and their standard errors from each regression. From Table 2, it follows that there is no correlation between the health outcomes and the skill measure.

Table 2: Covariate balance at hospital level

Covariate	**Coefficient**	**Standard Error**
Low birth weight	0.001	0.001
Prematurity	0.000	0.001
Apgar < 7	0.011	0.009
Antenatal consultations ≥ 4	0.000	0.003
Proportion of female newborns	0.000	0.001
Proportion of mothers with at least secundary education	-0.002	0.003
Proportion of married mothers	0.001	0.002
Proportion of teenage mothers	0.000	0.002
LBW $>$ p(75)	0.003	0.013
Prematurity $>$ p(75)	-0.004	0.011
Mean number of antenatal consultations	-0.005	0.022
Hospitals by municipalities	0.000	0.010
Municipality population	325.7	1,032.3

Notes: This table reports the results of regressing each hospital's characteristics. The data comes from the 2013-2016 DANE VSR, which collects and provides information that reveals the changes in mortality and fertility for each hospital. 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 1 is the proportion of newborns whose Apgar 1 score is lower than 7; Antenatal consultations ≥ 4 is the proportion of mothers who had at least for visits; Female newborn is the proportion of female newborns; married mothers is the proportion of married mothers; and teenage mothers is the proportion of mothers aged 19 years old or less. Calculations were made based on a sample of 2,365 physicians and 592 small hospitals. We interpret the non-significance of these estimates as evidence in favor of the randomness of the assignment of physicians.

5.2 Main results on health at birth

In this section, we provide our main results on health birth outcomes. Table 3 presents the estimated coefficient β, in equation (1), using ordinary least squares. We find that our main skill measure has a negative and significant effect on both low birth weight and the alternative measures of health.[28] The coefficient represents the effect of an increase of one standard deviation of physicians' average score. The standard error of the coefficient is presented in parenthesis, and we present the relative (percent) effect in square brackets—we divide the main coefficient by the average of the dependent variable.

In column 1 of Table 3, we see that there is a significant negative relationship between low birth weight and the average score—a decrease in the probability of being born low weight of 0.36 percentage points. Our estimates suggest that an increase of one standard deviation in physicians' average score decreases the probability of low birth weight by 7.42%.[29] Columns (2) and (3) in Table 3 examine alternative measures of health at birth. The point estimate for the standardized average score is associated with a decrease in the probability of being premature of 0.41 percentage points (10.05 %) and a drop in the probability of being born with an Apgar score below 7 of -0.63 percentage points (13.72 %). These results are consistent with previous literature that finds that prematurity is an important determinant of weight at birth (Almond et al., 2005).[30]

[28] All regressions include draw state and year of birth of the newborn fixed effects. The set of control variables includes: an indicator variable for the gender of the newborn, a 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 a teenager and zero otherwise, dummies for the mother's marital status, number of inhabitants in the municipality, number of hospitals per municipality, area of the municipality, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise.

[29] In the education context, the teacher value-added literature (e.g., Chetty et al., 2014; Rothstein, 2017) finds that an increase in teacher quality of one standard deviation corresponds 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 low birth weight by 7.4 percent. 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 the top-ranked program (see Figure 1).

[30] We find a strong correlation between prematurity and low birth weight in Colombia. Figure A.2 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 2010 and 2012. The figure presents the local polynomial regression fit of the probability of having a low birth weight over the number of gestational weeks using all birth records in Colombia from 2009 to 2012.

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

Table 3: Main estimates using all sample and average score

	Low birth weight	**Prematurity**	**Apgar < 7**
		Average Score	
	(1)	(2)	(3)
		With controls	
Coefficient	-0.0036	-0.0041	-0.0063
Standard Error	(0.0018)	(0.0015)	(0.0021)
Relative Effect	-7.42%	-10.05%	-13.72%
Average Dependent Variable	0.049	0.041	0.046
Number of Hospitals		592	
Number of Observations		104,357	

Notes: This table shows our main estimates. The coefficients represent the effect of an increase of one standard deviation of the physician skill measure (average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects and 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are 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.

5.2.1 Robustness Checks

We run additional specifications in which we use the standardized principal component instead of the standardized average score as a proxy for physicians' skills, using the full set of controls, and exclude the controls from our main estimating equation.[31] Figure 2 compares the estimated coefficient (relative to the mean), β, in equation (1) using the average score (main specification) with the three specifications mentioned above. We see from Figure 2 that our estimates for low birth weight are similar if we use the first principal component as a proxy for skills and are robust to the set of controls included in our analysis.[32]

Also, while ordinary least squares allows us to compute the average effect of our skills measure, it does not tell us much about the magnitude of this effect across the distribution of skills. We rank the skills into quartiles and estimate equation (1) using a set of dummies indicating the score distribution quartile to which physicians belonged. The results are presented in Appendix Table A.4. Columns (1), (3), and (5) present the coefficients associated with the effect of belonging to the second, third, and fourth quartile, respectively, of the average score distribution for low birth weight, relative to the first quartile. Columns (2), (4), and (6) present the coefficients associated with the effect of belonging to the second, third, and fourth quartile, respectively, of the distribution of the first principal component score on low birth weight 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, which suggests that there are potential gains associated with getting a more-skilled physician across the whole distribution of skills.

Finally, we extend our analysis by estimating additional models for low birth weight using alternative measures of skills. We aggregate health-related test scores (health management and public health) into a single *health score* and reading and quantitative test scores into a single *academic* score. We regress low birth weight on our health and academic scores, as well as on each individual exam score. Table 4 shows that the scores have a negative effect on low birth weight and are not statistically different from each other. The point estimates seem to be larger and more precisely estimated for the average health scores, especially for health management.[33]

[31] Note that the average prevalence of the outcomes considered is usually low and around 5%. 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 the skill measure. Appendix Table A.3 shows that the marginal effects (signs and magnitudes) are very similar to the ones estimated using a linear regression model.

[32] Results are reported in Appendix Table A.3, where we use low birth weight, prematurity, and Apgar score as our dependent variables, using the standardized average college examination score and standardized principal component as a proxy for physicians' skills.

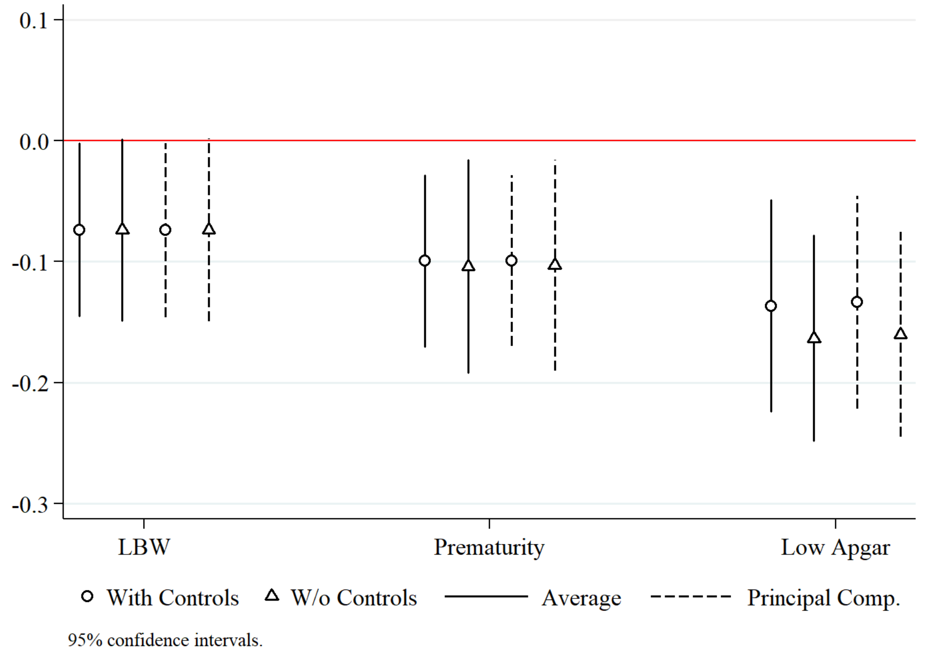
[33] The health module measures the physician's ability to assess primary health care. In particular, the health management score evaluates the physician's knowledge of the administrative processes necessary for the development of health activities and how well the physician would implement the patient safety and ethical standards necessary to provide health services. We use the SABER PRO scores in four areas (health

Table 4: Additional estimates using alternative measures of skills

	Health Score	Health Management Score	Public Health Score	Academic Score	Reading Score	Quantitative Score
	(1)	(2)	(3)	(4)	(5)	(6)
			With controls			
Coefficient	-0.0044	-0.0039	-0.0029	-0.0037	-0.0037	-0.0020
Standard Error	(0.0023)	(0.0018)	(0.0023)	(0.0021)	(0.0019)	(0.0017)
Relative Effect	-9.02%	-7.88%	-5.84%	-7.60%	-7.65%	-3.99%
Average Dependent Variable			0.049			
Number of Hospitals			592			
Number of Observations			104,357			

Notes: The coefficients in this table represent the effect of an increase of one standard deviation of the specific measure of physician skill. The Health score measure is the average of the Health Management and Public Health scores, and the Academic score is the average of the Reading and Quantitative scores. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state and year of birth of the newborn fixed effects and 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors. The table shows that while the coefficients are not statistically different from each other, the point estimates are consistently negative for all the scores and seem to be larger and more precisely estimated for the average Health scores.

Figure 2: Main estimates using all sample



Notes: The coefficients presented in this figure represent the relative effect of an increase of one standard deviation of the physician skill measure (average score or the first principal component of the four tests available). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity 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.

5.2.2 Placebo Tests

To evaluate our identification strategy's validity, we implement a placebo test, using VSR records for children born in 2010, 2011, and 2012. Recall that for our main results, we use

management, public health, reading, and quantitative test scores) as proxies of physicians' skills before the SSO program in our estimations. Other test scores, such as writing, citizenship skills, and English, are excluded since they were not tested in our sample cohorts.

data from the three years for which the doctors working at hospitals were randomly assigned (2013-2015). We move the physician's arrival time three years back and run placebo tests similar to our main specification but using data for the three previous years (2010-2012). We then estimate equation (1) using the same set of controls and fixed effects used in Table 3. Since physicians in our sample did not treat children born in 2010, 2011, and 2012, we would expect a null effect. Table 5 shows that the point estimates are precisely estimated zeros.[34] We see from Table A.5 that these results are similar if we use the first principal component as a proxy for skills and are robust to the set of controls included in our analysis.[35]

Table 5: Placebo test

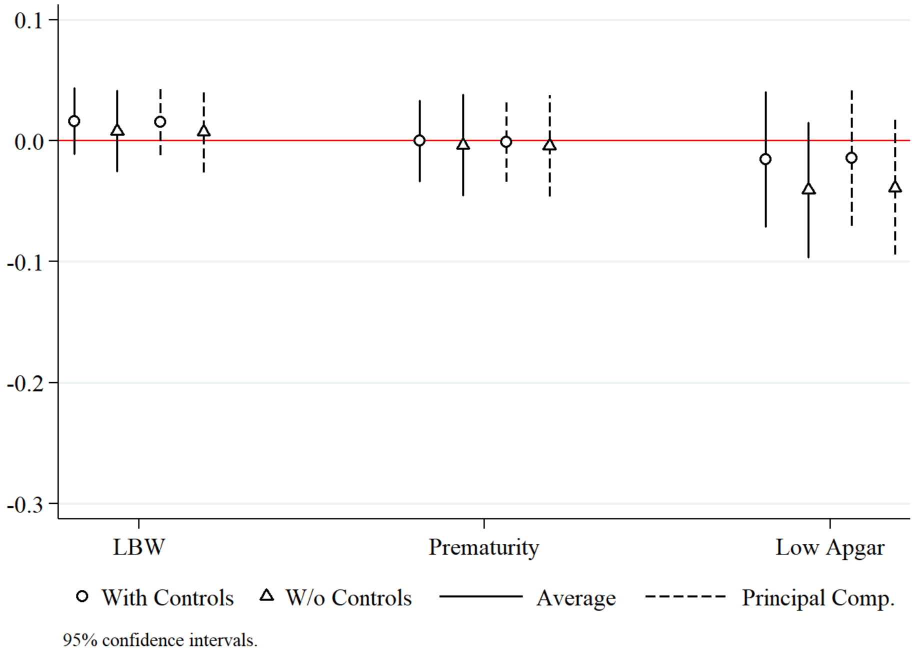
	Low birth weight	**Prematurity**	**Apgar < 7**
		Average Score	
	(1)	(2)	(3)
		With controls	
Coefficient	0.0012	0.0000	-0.0011
Standard Error	(0.0010)	(0.0010)	(0.0020)
Relative Effect	2.18%	0.01%	-2.05%
Average Dependent Variable	0.055	0.044	0.053
Number of Hospitals		600	
Number of Observations		102,050	

Notes: This table shows the results of running an exercise analogous to the one presented in Table 3 but moving the arrival date of the physician three years back (years 2010-2012). The coefficients represent the effect of an increase of one standard deviation of the physician skill measure (average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state and year of birth of the newborn fixed effects and 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are 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 hospitals.

[34] In Appendix Table A.6, we present the results for windows of 4, 3.5, 2.5, and 2 years before the start of the SSO program.

[35] For consistency, we implement the placebo test using an analogous Logit model and compute the average marginal effect associated with an increase in one standard deviation of the skill measure. Appendix Table A.7 shows that the marginal effects (signs and magnitudes) are null to the ones estimated using a linear regression model.

Figure 3: Placebo using all samples and average scores



Notes:.This figure shows the results of running an exercise analogous to the one presented in Figure 2 but moving the arrival date of the physician three years back (years 2010-2012). The coefficients represent the effect of an increase of one standard deviation of the physician skill measure (average score or the first principal component of the four tests available). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar score 1 of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. These results support the ones presented in Table 5 on the robustness of the estimated zero effect for the placebo tests.

5.3 Physicians' impacts across subgroups

In this section, we explore whether physicians' effects are more pronounced among some groups. Literature in economics has studied a variety of heterogeneous effects associated with socioeconomic status, measured by mother's education, age, and marital status (Amarante et al., 2016; Hoynes et al., 2011; Persson and Rossin-Slater, 2018) and gender of the newborn

(Almond and Mazumder, 2011; Currie and Schwandt, 2016a; Dinkelman, 2017; Eriksson et al., 2010; Okeke and Abubakar, 2020).[36] Although the effects are slightly more pronounced among less educated mothers, married women, first-time mothers, and non-teenage mothers (see Table 6), we do not find statistically significant differences on the effects across mothers' characteristics.

Similarly, we examine whether the treatment effects vary by the infant's gender. 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).[37] 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 low birth weight was particularly pronounced among male newborns, we do not find any statistical difference between males and females (see Table 7).

Finally, we look at heterogeneity across hospital characteristics. We divide the sample between hospitals below (low incidence) and above (high incidence) the 75th percentile of low birth weight distribution using data from the SSO program for the three years before our sample period (2010-2012). In Table 7, columns 1 and 2, we test the effects associated with physicians assigned to hospitals with a high incidence of low birth weight for these three years, which we interpret as hospitals with a high incidence of poor health outcomes (Currie, 2011). We do not find a significant effect of physicians' skills on low birth weight in hospitals with incidence of low birth weight. However, the effect is strongly negative and significant in hospitals with a high incidence of low birth weight. The point estimate for physicians in hospitals with high incidence is -0.73 percentage points (an increase of one standard deviation in physicians' average score decreases the probability of low birth weight by 11.66%), suggesting that physicians play a more important role in hospitals with a history of poor health outcomes.[38]

[36] Similar to other studies that focus on the VSR, our data includes information on the fetus's gender and mother's education, age, and marital status and whether she is a first-time mother.

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

[38] These results relate to the wide literature on heterogeneous clinical practices across hospitals and whether these differences translate into health outcomes. Doyle et al. (2015) find significant health benefits for older patients who are brought to higher-cost hospitals, Card et al. (2019) finds that, during the first year of life, newborns who were delivered by c-sections are more likely to visit the emergency department, less likely to be readmitted to hospital, and have lower mortality rates. Related contributions include Cutler et al. (2019) and Finkelstein et al. (2016). See Skinner (2011) for a review of the literature on regional variation in intensity of care or spending.

Table 6: Heterogeneity of the effects across mothers' characteristics

	Low birth weight							
	Mother with low education	Mother with high education	Married mother	Single mother	First-time	Non-first-time	Teenage mother	Non-teenage mother
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Score average							
Coefficient	-0.0030	-0.0026	-0.0057	-0.0011	-0.0036	-0.0021	-0.0024	-0.0036
Standard Error	(0.0017)	(0.0023)	(0.0017)	(0.0019)	(0.0018)	(0.0015)	(0.0017)	(0.0019)
Relative Effect	-6.03%	-5.65%	-13.33%	-2.02%	-7.42%	-5.58%	-4.46%	-8.24%
Average Dependent Variable	0.050	0.046	0.043	0.053	0.049	0.038	0.054	0.044
Number of Observations	89,599	14,751	39,921	64,430	104,357	152,447	57,076	47,276

Notes: This table shows the heterogeneity of our estimated results when we divide the sample by mothers' characteristics. The coefficients represent the effect of an increase of one standard deviation of the physician skill measure (score average) for each subgroup. Relative (percent) effects are in square brackets and are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. A mother is considered to be high (low) education when she has any (no) level of tertiary education. A teenage mother is someone who has given birth at age 19 years old or younger. All regressions control for draw state and year of birth of the newborn fixed effects and 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors. We interpret these results as lack of evidence of any statistically significant difference in the effects across the observed mothers' characteristics.

6 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 et al., 2017). Some of these practices, like the quality of medical advice provided by doctors, are unobservable (Das et al., 2008; Leonard and Masatu, 2007), whereas others, like the number of prenatal consultations, are observable. In this section, we study potential mechanisms for observed differences between skilled and unskilled physicians.

6.1 Full exposure to the treatment

First, we study the impact of the overlap between the one-year service period of the SSO program and women's gestation period. We split the sample into mothers for whom the entire gestation period is covered in the physicians' one-year service period (fully exposed) and the complementary group (partially exposed), which includes mothers who are only partially exposed. We find that the effect is negative and statistically significant only for cases in which physicians had full continuity of care (see Table 7). The point estimate for fully exposed mothers is -0.41 percentage points (an increase of one standard deviation in physicians' average score decreases the probability of low birth weight by 7.14%), while for mothers who were partially exposed, the effect is -0.31 percentage points (decrease in probability of low birth weight of 6.5%) but it is less precisely estimated. The potential time each mother spends with the doctor during pregnancy seems to be an important driver of the main effects.

Table 7: Heterogeneity of the effects across hospital and pregnancy characteristics

	Low birth weight					
	Hospital		Pregnancy			
	Higher incidence of LBW	Lower incidence of LBW	Female newborns	Male newborns	Full Continuity of care	Partial Continuity of care
	(1)	(2)	(3)	(4)	(5)	(6)
			Average score			
Coefficient	-0.0073	-0.0004	-0.0007	-0.0048	-0.0041	-0.0031
Standard Error	(0.0025)	(0.0017)	(0.0016)	(0.0021)	(0.0020)	(0.0021)
Adjusted Coefficient	-11.66%	-1.02%	-1.22%	-10.65%	-7.14%	-6.50%
Average Dependent Variable	0.063	0.039	0.054	0.045	0.057	0.047
Number of Observations	46,292	58,060	50,820	53,534	26,862	77,487

Notes: The coefficients represent the effect of an increase of one standard deviation of the physician skill measure (average score). Relative (percent) effects are in square brackets and are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar score 1 of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects. In panel A, estimations were made without additional control variables, while in Panel B, we 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors. In columns 1 and 2, we divide the sample between hospitals below (low incidence) and above (high incidence) the 75th percentile of low birth weight distribution using data from the SSO program for the three years before our sample period (2010-2012). The results suggest that physicians play a more important role in hospitals with a history of poor health outcomes. Although we find that the reduction in low birth weight was particularly pronounced among male newborns, we do not find any statistical difference between males and females. In columns 5 and 6, we split the sample into mothers for whom the entire gestation period is covered in the physicians' one-year service period (fully exposed) and the complementary group (partially exposed), which includes mothers who are only partially exposed. We find that the effect seems to be stronger and more precisely estimated for fully exposed mothers compared to partially exposed mothers.

6.2 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. According to WHO (2016) and Colombian government (Gomez et al., 2013), prenatal care improves the health status of both mother and newborn. In Colombia, the Ministry of Health requires the prenatal monitoring be carried out by physicians (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 scheduled mothers to have four or more prenatal checkups (see Table A.8). 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; Carrillo and Feres, 2019; Conway and Deb, 2005; Currie and Grogger, 2002; Grossman and Joyce, 1990; Kramer, 1987; McCormick and Siegel, 2001).[39]

We expect that physicians enrolled in the SSO program and assigned to rural areas of our sample (outside the metropolitan areas) would be time constrained, as usually they are the only physicians available in those areas. This is supported by anecdotal evidence described in different reports from Colombian medical associations in which physicians refer to the SSO year as an experience where they had an overwhelming workload and long working hours.[40] In this setting where physicians are time constrained, it comes as no surprise that the average likelihood of having sufficient prenatal consultations remains unaffected 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 likely to give birth to a baby with low birth weight.

We assume that low birth weight can be thought of as a prediction problem, and take advantage of recent advances in machine learning techniques.[41] We use several of these techniques to generate two groups of predictions about mothers' probability of low birth weight using a set of mother-hospital 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, XGBoost, neural networks, and logistic regression models.[42]

[39] Barber and Gertler (2010), exploit the random initial assignment of the Mexican *Progresa/Oportunidades* and find a large reduction in the incidence of low birth weight, which they attribute to better-quality prenatal care.

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

[41] Supervised machine learning seeks to solve the problem of prediction (Kleinberg et al., 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.

[42] 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

We train each of these algorithms on a sample of children born in a set of randomly selected hospitals representing 25% of the total number of hospitals in our main sample. We follow Chernozhukov et al. (2018) and re-scale the outcomes and covariates to be between 0 and 1 before training in all the machine learning methods.

We fit the model to the training sample with the four different methods and predict on the test sample. We then divide the test sample into two groups: low and high predicted probability, defined as mothers with a probability of low birth weight below and above the median, respectively, for each of the four predictions.[43] We then estimate equation 1 using a dummy that is equal to 1 if the number of prenatal consultations is four or above as our main outcome in each of the groups defined before (i.e., low and high predicted probability of low birth weight). The results of these regressions are presented in Table 8. Columns (1) and (2) present the results for the sample of mothers with a low predicted probability, and columns (3) and (4) present the results for the sample of mothers with a high predicted probability of low birth weight. We include both the average college examination score score and the principal component average as the measure of physician skills.

Table 8 shows that regardless of the method we use to predict low birth weight, more-skilled doctors do not seem to increase the recommended number of antenatal consultations for mothers with a low predicted probability of low birth weight. Instead, they target those prenatal checkups toward the more vulnerable mothers, measured as mothers with a higher predicted probability of low birth weight. Consistent with our suggested mechanism of physicians being able to target care toward the more vulnerable mothers, we find stronger effects of our measure of skills when we focus on mothers with a higher predicted probability of low birth weight compared to those with lower predicted probability. While the point estimate for the effect of physicians' test score on low birth weight in the lower predicted probability sample is between 0.16 and 0.59 percentage points depending on the prediction used to divide the data, the point estimate for the higher predicted probability group is between 1.34 and 2.2 percentage points. Altogether, the results from this section are consistent with a story of time-constrained physicians not being able to increase the average amount of time spent in prenatal consultations but improving the targeting of care toward the more vulnerable mothers.

We next show that, consistent with the idea of better physicians being better at targeting

of algorithms in specific settings. See Mullainathan and Spiess (2017) for a review of the literature and Breiman (2001) for a description of the process of the methods.

[43] A similar strategy is followed by Liberman et al. (2018) and Liberman et al. (2021), who study the effects of information deletion and usury rates on consumer credit markets.

Table 8: Antenatal consultations by predicted low birth weight

	Dependent variable: Antenatal consultations ≥ 4			
	Low predicted probability of low birth weight		**High predicted probability of low birth weight**	
	Score average (1)	PCA score (2)	Score average (3)	PCA score (4)
		With controls		
		Panel A. Logistic regression model		
Coefficient	0.0026	0.0030	0.0172	0.0175
Standard Error	(0.0050)	(0.0051)	(0.0061)	(0.0061)
Relative Effect	0.29%	0.33%	2.04%	2.08%
		Panel B. Random forest		
Coefficient	0.0054	0.0059	0.0134	0.0136
Standard Error	(0.0049)	(0.0050)	(0.0060)	(0.0060)
Relative Effect	0.61%	0.67%	1.57%	1.60%
		Panel C. XGBoost		
Coefficient	0.0034	0.0037	0.0159	0.0164
Standard Error	(0.0048)	(0.0048)	(0.0060)	(0.0060)
Relative Effect	0.38%	0.42%	1.89%	1.94%
		Panel D. Neural networks		
Coefficient	-0.0020	-0.0016	0.0220	0.0222
Standard Error	(0.0052)	(0.0053)	(0.0070)	(0.0070)
Relative Effect	-0.22%	-0.18%	2.60%	2.62%

Notes: This table reports the differential effects of the skill measure on antenatal consultations by mother's predicted probability of low birth weight. To predict the probability of low birth weight, we train four different types of models (random forest, XGBoost, neural networks, and logistic regression models) on a sample of children born in a set of randomly selected hospitals representing 25% of the total number of hospitals in our main sample. We then predict for the remaining sample and divide it into two groups: low and high predicted probability, defined as mothers with a probability of low birth weight below and above the median, respectively, for each of the four predictions. The coefficients presented represent the effect of an increase of one standard deviation of the physician skill measure (average score or the first principal component of the four tests available) on the probability of having four or more antenatal consultations. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. All regressions control for draw state and year of birth of the newborn fixed effects. All estimations include the controls from the main estimation (see Table 3). Numbers in parentheses are clustered standard errors. The results show that regardless of the method we use to predict low birth weight, more-skilled doctors do not seem to increase the recommended number of antenatal consultations for mothers with a low predicted probability of low birth weight. Instead, they target those prenatal checkups toward the more vulnerable mothers, measured as mothers with a higher predicted probability of low birth weight.

care to the most vulnerable mothers, the negative effects on the probability of having low birth weight, prematurity, or low Apgar score are particularly pronounced among the more vulnerable mothers. Table 9 shows that regardless of the method we use to split the sample, more-skilled doctors do seem to improve health at birth mainly for mothers with (ex ante) high predicted probability of low birth weight. In particular, for the more vulnerable mothers, an increase of one standard deviation in physicians' average college examination score decreases the probability of low birth weight between 7.14% and 10.6%, the probability of prematurity between 11.55% and 15.15%, and the probability of low Apgar between 18.05% and 19.36%, while for mothers with (ex ante) low predicted probability of low birth weight, the effects are smaller in magnitude and not statistically different from zero.

Table 9: Main outcomes by predicted low birth weight

	Low Birth Weight		**Prematurity**		**Apgar < 7**	
	Low Predicted Low Birth Weight (1)	High Predicted Low Birth Weight (2)	Low Predicted Low Birth Weight (3)	High Predicted Low Birth Weight (4)	Low Predicted Low Birth Weight (5)	High Predicted Low Birth Weight (6)
With controls						
Panel A. Logistic regression model						
Coefficient	-0.0020	-0.0047	-0.0008	-0.0057	-0.0060	-0.0087
Standard Error	(0.0015)	(0.0025)	(0.0019)	(0.0022)	(0.0037)	(0.0043)
Relative Effect	-4.99%	-7.96%	-2.43%	-11.55%	-13.96%	-18.05%
Panel B. Random forest						
Coefficient	-0.0014	-0.0059	-0.0002	-0.0071	-0.0056	-0.0085
Standard Error	(0.0015)	(0.0020)	(0.0017)	(0.0020)	(0.0035)	(0.0036)
Relative Effect	-3.19%	-10.55%	-0.44%	-15.15%	-12.18%	-18.78%
Panel C. XGBoost						
Coefficient	-0.0022	-0.0042	-0.0020	-0.0057	-0.0049	-0.0091
Standard Error	(0.0015)	(0.0021)	(0.0018)	(0.0021)	(0.0035)	(0.0035)
Relative effect	-5.61%	-7.14%	-6.11%	-11.69%	-11.11%	-19.36%
Panel D. Neural networks						
Coefficient	-0.0020	-0.0047	-0.0008	-0.0057	-0.0060	-0.0087
Standard Error	(0.0015)	(0.0025)	(0.0019)	(0.0022)	(0.0037)	(0.0043)
Relative Effect	-4.99%	-7.96%	-2.43%	-11.55%	-13.96%	-18.05%

Notes: This table reports the differential effects of physicians' skill measure on main outcomes by mother's predicted probability of low birth weight. We divide the sample as in 8. The coefficients represent the effect of an increase of one standard deviation of the physician skill measure (average score or the first principal component of the four tests available). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. All regressions control for draw state and year of birth of the newborn fixed effects. All estimations include the controls from the main estimation (see Table 3). Numbers in parentheses are 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 having low birth weight, prematurity, or low Apgar score are particularly pronounced among the more vulnerable mothers.

7 Conclusions

Physicians are a key input in the production of 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 hospitals (Doyle et al., 2010). In this 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, hospitals, vital statistics records, characteristics of the physicians, and the mandatory field-specific college graduation exams to measure the skills of the physicians assigned to each hospital and the main health outcomes. Using these datasets, we provide evidence of the covariate balance between winners and losers of the SSO program, and between hospitals and the quality of physicians. Finally, we provide evidence of the causal relationship between more-skilled physicians and health at birth.

We find that more-skilled physicians have a negative and significant effect on the probability of low birth weight. We estimate that an increase in one standard deviation in the physicians' academic test score reduces the probability of low birth weight by 7.4%. Although low birth weight is our main measure of health at birth, the results are robust to other measures such as prematurity and Apgar score. Second, we document that these effects are entirely driven by hospitals with high incidence of poor health of newborn the years before the program.

Further, we explore the importance of the mother's contact with the physician as a potential mechanism that mediates this impact. We focus on two mechanisms: First, we look at the potential time each mother spends with the doctor during pregnancy. We find much stronger effects in cases where the mother has more exposure time with the physician. Second, we explore whether more-skilled physicians increase 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 to have 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 a baby with low birth weight.

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 in in utero health and that these findings should be considered by governments in developing policies to assign physicians optimally.

References

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

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

Adhvaryu, A., Nyshadham, A., Molina, T., and Tamayo, J. (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 Korenbrot, C. C. (1995). The role of prenatal care in preventing low birth weight. *The Future of Children*, pages 103–120.

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

Almond, D., Currie, J., and Duque, V. (2018). Childhood circumstances and adult outcomes: Act ii. *Journal of Economic Literature*, 56(4):1360–1446.

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

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

Amarante, V., Manacorda, M., Miguel, E., and Vigorito, A. (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., Dobkin, C., and Gross, T. (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., Dobkin, C., and Gross, T. (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.

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

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

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

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

Barber, S. L. and Gertler, P. J. (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., Wang, J. J., De Leon, S. F., Shih, S. C., Boscardin, W. J., Goldman, L. E., and Dudley, R. A. (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., Gertler, P. J., Binagwaho, A., Soucat, A. L., Sturdy, J., and Vermeersch, C. M. (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.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2007). From the cradle to the labor market? the effect of birth weight on adult outcomes. *The Quarterly Journal of Economics*, 122(1):409–439.

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

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

Callaghan, W. M., MacDorman, M. F., Rasmussen, S. A., Qin, C., and Lackritz, E. M. (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., Fenizia, A., and Silver, D. (2019). The health impacts of hospital delivery practices. Technical report, National Bureau of Economic Research.

Carrera, M., Goldman, D. P., Joyce, G., and Sood, N. (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 Feres, J. (2019). Provider supply, utilization, and infant health: evidence from a physician distribution policy. *American Economic Journal: Economic Policy*, 11(3):156–96.

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

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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

Chetty, R., Friedman, J., and Rockoff, J. (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

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

Clemens, J. and Gottlieb, J. D. (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.

Conway, K. S. and Deb, P. (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 Almond, D. (2011). Human capital development before age five. In *Handbook of Labor Economics*, volume 4, pages 1315–1486. Elsevier.

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

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

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

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

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

Cutler, D., Skinner, J. S., Stern, A. D., and Wennberg, D. (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 Hammer, J. (2005). Which doctor? combining vignettes and item response to measure clinical competence. *Journal of Development Economics*, 78(2):348–383.

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

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

Das, J., Holla, A., Mohpal, A., and Muralidharan, K. (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 Sohnesen, T. P. (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., Ewer, S. M., and Wagner, T. H. (2010). Returns to physician human capital: Evidence from patients randomized to physician teams. *Journal of Health Economics*, 29(6):866–882.

Doyle, J. J., Graves, J. A., Gruber, J., and Kleiner, S. A. (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., Kajantie, E., Osmond, C., Thornburg, K., and Barker, D. J. (2010). Boys live dangerously in the womb. *American Journal of Human Biology*, 22(3):330–335.

Fernández Ávila, D. G., Mancipe García, L. C., Fernández Ávila, D. C., Reyes Sanmiguel, E., Díaz, M. C., and Gutiérrez, J. M. (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., Gentzkow, M., and Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics*, 131(4):1681–1726.

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

Gagnon-Bartsch, J., Shem-Tov, Y., 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., Arevalo, I., 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 Gilleskie, D. (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 Joyce, T. J. (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., Posso, C., Saravia, E., and Tamayo, J. (2021). Healing the gender gap: The impacts of randomized first-job on female physicians.

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

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

Hoffman, M., Kahn, L. B., and Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800.

Hoynes, H., Page, M., and Stevens, A. H. (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.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (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 Masatu, M. C. (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., Masatu, M. C., and Vialou, A. (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., Medina, C., Neilson, C., and Posso, C. (2021). Lender market power and the bright side of interest rate caps: Evidence from colombia. Technical report, Unpublished manuscript.

Liberman, A., Neilson, C., Opazo, L., and Zimmerman, S. (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.

Ma, S. and Finch, B. K. (2010). Birth outcome measures and infant mortality. *Population Research and Policy Review*, 29(6):865–891.

Mathews, F., Johnson, P. J., and Neil, A. (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 Siegel, J. E. (2001). Recent evidence on the effectiveness of prenatal care. *Ambulatory Pediatrics*, 1(6):321–325.

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

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

Moster, D., Lie, R., and Markestad, T. (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 Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

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

Okeke, E. N. and Abubakar, I. S. (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., Stabile, M., Walld, R., and Roos, L. L. (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., Jaramillo, L., Franco, C., and Arregoces, L. (2007). Estudio sobre el modo de gestionar la salud en colombia.

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

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

Razaz, N., Boyce, W. T., Brownell, M., Jutte, D., Tremlett, H., Marrie, R. A., and Joseph, K. (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., Hanushek, E. A., and Kain, J. F. (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 Currie, J. (2018). Addressing the opioid epidemic: is there a role for physician education? *American Journal of Health Economics*, 4(3):383–410.

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

Simeonova, E., Skipper, N., and Thingholm, P. R. (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, pages 45–93. Elsevier.

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

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

Tsugawa, Y., Jena, A. B., Figueroa, J. F., Orav, E. J., Blumenthal, D. M., and Jha, A. K. (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., Kenny, D. T., Gibson, F., Bowen, J., and Starte, D. (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.

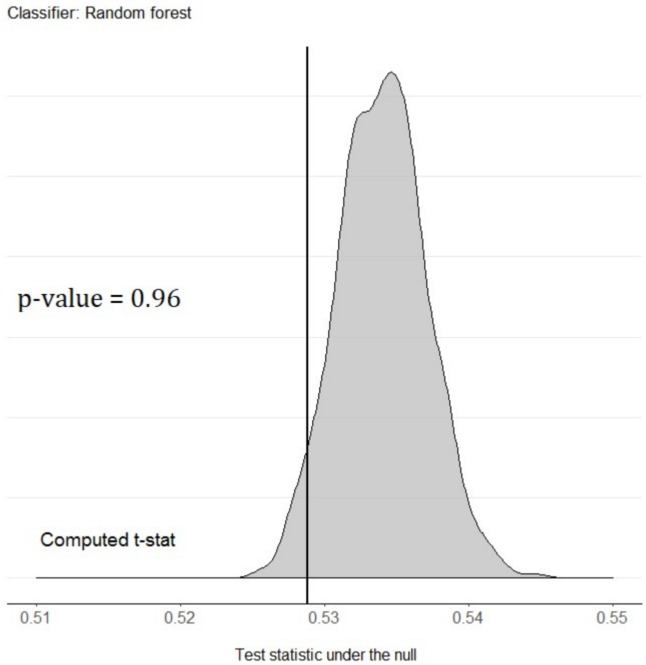
Wenghofer, E., Klass, D., Abrahamowicz, M., Dauphinee, D., Jacques, A., Smee, S., Blackmore, D., Winslade, N., Reidel, K., Bartman, I., et al. (2009). Doctor scores on national qualifying examinations predict quality of care in future practice. *Medical education*, 43(12):1166–1173.

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.

A Appendix

Figure A.1: Balancing test using the Classification Permutation Test (Gagnon-Bartsch and Shem-Tov, 2018)



Notes: This graph shows the results for the Classification Permutation Test: A Machine Learning Nonparametric Test for Equality of Multivariate Distributions (Johann Gagnon-Bartsch and Yotam Shem-Tov, 2018, *Annals of Applied Statistics*). The procedure includes 1,000 repetitions. We also perform a reverse regression test (F(19, 160) = 0.88, P-value = 0.6128). These result provide additional evidence in favor of the randomization

Table A.1: Balancing rural winners and losers

Covariable	**Control Mean**	**Standard Deviation**	**Coefficient**	**Standard Error**
The household has a private car	0.497	0.500	0.011	0.019
Gender (female)	0.590	0.492	-0.008	0.021
Number of people in the household	3.960	1.650	0.038	0.048
Father with tertiary education	0.667	0.471	-0.009	0.018
Mother with tertiary education	0.669	0.471	-0.012	0.015
Socioeconomic strata: 1 or 2 or rural areas	0.219	0.414	0.024	0.017
Socioeconomic strata: 4, 5, or 6	0.425	0.494	-0.009	0.015
Level of SISBEN: 1 or 2	0.219	0.414	0.008	0.017
The household has internet	0.868	0.339	-0.006	0.012
Monthly household income: Less than 2 MW	0.211	0.408	0.003	0.016
Monthly household income: $\geq$ 2 and ¡ 3 MW	0.199	0.399	0.008	0.014
The father or the mother has a job	0.877	0.328	0.002	0.015
The household has a washing machine	0.878	0.328	0.005	0.009
The household has a television	0.870	0.336	0.013	0.011
The household has a cellphone	0.968	0.177	-0.003	0.008
The house has proper flooring	0.936	0.245	-0.010	0.009
The household has an oven	0.718	0.450	-0.005	0.016
Physician's score on the reading test (ECAES)	10.688	0.966	-0.015	0.034
Physician's score on the Health Management test (ECAES)	10.419	1.036	0.011	0.032
Physician's average score on ECAES 4	10.539	0.833	0.007	0.028

Notes: This table reports lottery losers' means and estimated effects of winning the SSO, based on a sample of 3,559 observations with a 3,519-degree of freedom, testing a total of 20 hypotheses. Standard errors are clustered, given the by draw and state design of the randomization. Controls for draw-by-state fixed effects are included in the model. The eligibility for the subsidized regime is defined by the SISBEN score. SISBEN levels 1 and 2 are associated with the highest level of prioritization.

Table A.2: Main estimates using all sample

	Low Birth Weight		**Prematurity**		**Apgar < 7**	
	Average Score (1)	PCA Score (2)	Average Score (3)	PCA score (4)	Average Score (5)	PCA Score (6)
	Panel A. Without controls					
Coefficient	-0.0036	-0.0036	-0.0043	-0.0043	-0.0076	-0.0074
Standard Error	(0.0019)	(0.0019)	(0.0019)	(0.0018)	(0.0020)	(0.0020)
Relative Effect	-7.45%	-7.44%	-10.51%	-10.42%	-16.42%	-16.09%
	Panel B. With controls					
Coefficient	-0.0036	-0.0037	-0.0041	-0.0041	-0.0063	-0.0062
Standard Error	(0.0018)	(0.0018)	(0.0015)	(0.0015)	(0.0021)	(0.0021)
Relative Effect	-7.42%	-7.47%	-10.05%	-10.01%	-13.72%	-13.43%
Average Dependent Variable	0.049	0.049	0.041	0.041	0.046	0.046
S.D. Dependent Variable	0.217	0.217	0.199	0.199	0.210	0.210
Number of Hospitals	592	592	592	592	592	592
Number of Observations	104,357	104,357	104,357	104,357	104,357	104,357

Notes: The coefficients represent the effect of an increase of one standard deviation of the physician skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects. In panel A, estimations were made without additional control variables, while in Panel B, we 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors.

Table A.3: Main estimates logit robustness check

	Low Birth Weight		**Prematurity**		**Apgar < 7**	
	Average Score (1)	PCA Score (2)	Average Score (3)	PCA Score (4)	Average Score (5)	PCA Score (6)
			Panel A. Without controls			
Coefficient	-0.0035	-0.0035	-0.0045	-0.0044	-0.0072	-0.0071
Standard Error	(0.0017)	(0.0017)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Relative Effect	-6.95%	-6.90%	-10.65%	-10.48%	-15.76%	-15.49%
			Panel B. With controls			
Coefficient	-0.0034	-0.0034	-0.0042	-0.0042	-0.0058	-0.0057
Standard Error	(0.0016)	(0.0016)	(0.0013)	(0.0013)	(0.0019)	(0.0019)
Relative Effect	-6.84%	-6.86%	-10.00%	-9.92%	-12.64%	-12.44%
Average Dependent Variable	0.050	0.050	0.042	0.042	0.046	0.046
S.D. Dependent Variable	0.217	0.217	0.200	0.200	0.210	0.210
Number of Hospitals	577	577	579	579	586	586
Number of Observations	104,106	104,106	103,944	103,944	104,184	104,184

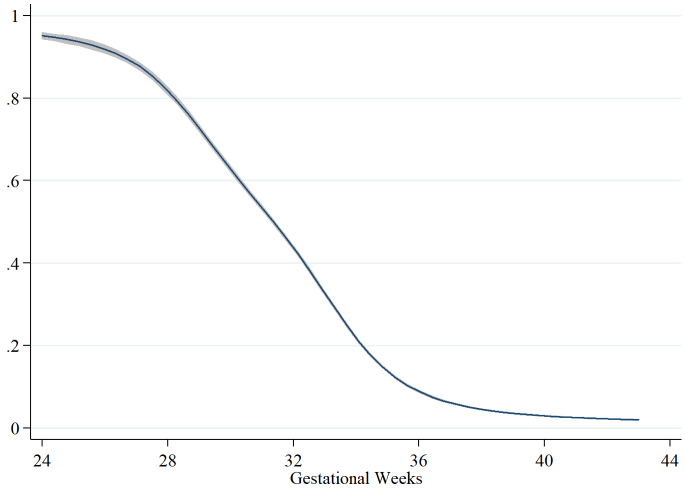
Notes: The coefficients represent the average marginal effect of the physician skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects. In panel A, estimations were made without additional control variables, while in Panel B, we 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors. The results are very similar to the ones found using linear regressions in our main estimates.

Table A.4: Main estimates linearity

	Low Birth Weight					
	Quartile 2		**Quartile 3**		**Quartile 4**	
	Average Score (1)	PCA Score (2)	Average Score (3)	PCA Score (4)	Average Score (5)	PCA Score (6)
	With controls					
Coefficient	-0.0031	-0.0033	-0.0042	-0.0052	-0.0054	-0.0065
Standard Error	(0.0036)	(0.0036)	(0.0038)	(0.0037)	(0.0042)	(0.0041)
Relative Effect	-6.24%	-6.66%	-8.65%	-10.70%	-11.11%	-13.19%
Average Dependent Variable	0.049	0.049	0.049	0.049	0.049	0.049
S.D. Dependent Variable	0.217	0.217	0.217	0.217	0.217	0.217
Number of Hospitals	592	592	592	592	592	592
Number of Observations	104,357	104,357	104,357	104,357	104,357	104,357

Notes: The coefficients represent the effect of being assigned a physician of the quartiles 2 (columns 1 and 2), 3 (columns 3 and 4), or 4 (columns 5 and 6) of the distribution of skills 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. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects. In panel A, estimations were made without additional control variables, while in Panel B, we 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 one if the hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors. While the coefficients are not statistically different, we do observe increases in the point estimates associated with higher quartiles and cannot discard linearity of the effects.

Figure A.2: 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.

Table A.5: Placebo three years robustness checks

	Low Birth Weight		**Prematurity**		**Apgar < 7**	
	Average Score (1)	PCA Score (2)	Average Score (3)	PCA Score (4)	Average Score (5)	PCA Score (6)
			Panel A. Without controls			
Coefficient	0.0006	0.0005	-0.0002	-0.0003	-0.0028	-0.0027
Standard Error	(0.0013)	(0.0013)	(0.0012)	(0.0013)	(0.0020)	(0.0020)
Relative Effect	1.05%	0.98%	-0.49%	-0.57%	-5.36%	-5.12%
			Panel B. With controls			
Coefficient	0.0012	0.0011	0.000002	-0.00004	-0.0011	-0.0010
Standard Error	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0020)	(0.0020)
Relative Effect	2.18%	2.05%	0.01%	-0.10%	-2.05%	-1.92%
Average Dependent Variable	0.055	0.055	0.044	0.044	0.053	0.053
S.D. Dependent Variable	0.229	0.229	0.206	0.206	0.224	0.224
Number of Hospitals	600	600	600	600	600	600
Number of Observations	102,050	102,050	102,050	102,050	102,050	102,050

Notes: The coefficients represent the effect of an increase of one standard deviation of the physician skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects. In panel A, estimations were made without additional control variables, while in Panel B, we 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 one if the hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors.

Table A.6: Placebo other years

	Low birth weight		**Prematurity**		**Apgar < 7**	
	Average Score	PCA score	Average Score	PCA score	Average Score	PCA score
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A. Placebo 2 years			
Coefficient	-0.0004	-0.0004	-0.0002	-0.0002	-0.0020	-0.0019
Standard Error	(0.0010)	(0.0010)	(0.0011)	(0.0011)	(0.0021)	(0.0021)
Relative Effect	-0.74%	-0.79%	-0.55%	-0.42%	-4.10%	-3.80%
			Panel B. Placebo 2.5 years			
Coefficient	0.0012	0.0012	-0.0009	-0.0009	-0.0024	-0.0023
Standard Error	(0.0009)	(0.0009)	(0.0013)	(0.0013)	(0.0020)	(0.0020)
Relative Effect	2.29%	2.26%	-2.20%	-2.11%	-4.82%	-4.60%
			Panel C. Placebo 3.5 years			
Coefficient	0.0012	0.0011	0.0000	0.0000	-0.0011	-0.0010
Standard Error	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0020)	(0.0020)
Relative Effect	2.17%	2.03%	0.00%	-0.10%	-2.05%	-1.92%
			Panel D. Placebo 4 years			
Coefficient	0.0014	0.0013	-0.0012	-0.0012	-0.0006	-0.0005
Standard Error	(0.0010)	(0.0010)	(0.0011)	(0.0011)	(0.0018)	(0.0018)
Relative Effect	2.57%	2.45%	-2.61%	-2.69%	-0.67%	-0.55%

Notes: The coefficients represent the effect of an increase of one standard deviation of the physician skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects and 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors.

Table A.7: Placebo logistic regression model

	Low Birth Weight		**Prematurity**		**Apgar < 7**	
	Average Score (1)	PCA Score (2)	Average Score (3)	PCA Score (4)	Average Score (5)	PCA Score (6)
			Panel A. Without controls			
Coefficient	0.0006	0.0006	-0.0002	-0.0002	-0.0030	-0.0029
Standard Error	(0.0013)	(0.0013)	(0.0014)	(0.0014)	(0.0021)	(0.0021)
Relative Effect	1.10%	1.03%	-0.46%	-0.54%	-5.72%	-5.49%
			Panel B. With controls			
Coefficient	0.0015	0.0014	0.0007	0.0007	-0.0012	-0.0011
Standard Error	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0021)	(0.0021)
Relative effect	2.62%	2.49%	1.60%	1.47%	-2.19%	-2.10%
Average Dependent Variable	0.056	0.056	0.045	0.045	0.053	0.053
S.D. Dependent Variable	0.229	0.229	0.206	0.206	0.224	0.224
Number of Hospitals	589	589	587	587	594	594
Number of Observations	101,557	101,557	101,510	101,510	101,944	101,944

Notes: The coefficients represent the effect of an increase of one standard deviation of the physician skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar score 1 of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects. In panel A, estimations were made without additional control variables, while in Panel B, we 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 hospitals per municipality, area, a dummy variable that indicates if there is at least one female physician in the cohort, a dummy variable that indicates if there is at least one physician from a top university in the cohort, a dummy variable that indicates if there is at least one physician from a public university in the cohort, an indicator variable that takes the value of 1 if the hospital 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 hospital is above the 75th percentile of the distribution of prematurity measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors.

Table A.8: Antenatal consultations

	Dependent Variable: Antenatal Consultations $\geq$ 4	
	Average Score	PCA Score
	(1)	(2)
	Panel A. Without controls	
Coefficient	-0.0004	-0.0003
Standard Error	(0.0072)	(0.0073)
Relative Effect	-0.05%	-0.04%
	Panel B. With controls	
Coefficient	-0.0017	-0.0016
Standard Error	(0.0069)	(0.0070)
Relative Effect	-0.20%	-0.19%
Average Dependent Variable	0.867	0.867
S.D. Dependent Variable	0.340	0.340
Number of Hospitals	592	592
Number of Observations	104,357	104,357

Notes: The coefficients represent the effect of an increase of one standard deviation of the physician skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. 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, and Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state and year of birth of the newborn fixed effects. Numbers in parentheses are clustered standard errors.