HARVARD | BUSINESS | SCHOOL

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

Arlen Guarin
Christian Posso
Estefania Saravia
Jorge Tamayo

Working Paper 22-015

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

Arlen Guarin
University of California, Berkeley

Christian Posso
Banco de la Republica de Colombia

Estefania Saravia
Harvard Business School

Jorge Tamayo
Harvard Business School

**Working Paper 22-015**



Funding for this research was provided in part by Harvard Business School and UC Berkeley Opportunity Lab.

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

Arlen Guarin, Christian Posso, Estefania Saravia, Jorge Tamayo*

November 2, 2021

**Abstract**

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

**Keywords:** Physicians' health 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, Brayan Pineda, and Carolina Velez for excellent research assistance. We thank Maria Aristizabal, Carolina Arteaga, Francesco Bogliacino, Leonardo Bonilla, David Card, Maíra Coube, Janet Currie, Kaveh Danesh, Margarita Gafaro, Robert Gonzalez, Hilary Hoynes, Rembrand Koning, Juliana Londoño-Vélez, Edward Miguel, Paul Rodriguez, Emmanuel Saez, Molly Schnell, Christopher Walters, Danny Yagan and seminar participants at Banco de la Republica, NBER SI Children, ESSEN health conference 2020, EEA congress 2020, Universidad Eafit, RIDGE LACEA's Health Economics Network, Universidad del Rosario, UC Berkeley (Development and Labor Lunch) for insightful comments. Arlen gratefully acknowledges financial support from the UC Berkeley Opportunity Lab. The findings, interpretations, and conclusions expressed in this paper do not necessarily reflect the views of Banco de la República or its Board of Directors. 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 predict future abilities and health trajectories that genetics cannot explain (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). Notably, an unresolved important question is whether more-skilled healthcare professionals can improve health outcomes at birth.

In this paper, we break 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 (Chan Jr et al., 2019; Chen, 2021; Currie and MacLeod, 2017, 2020; 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, because answering this question poses a substantial empirical challenge. It requires accounting 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 national government program that randomly assigned 2,126 physicians to 618 small hospitals. We estimate the impact on the 256,806 children whose mothers received care in those hospitals during their pregnancy, using administrative data from the program, 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 example, 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] To measure the medical skills of recent medical graduates, we use physician's individual records from the country's mandatory, field-specific college graduation exams. 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. A key feature in our setting is that hospitals' characteristics do not covariate with physicians' skills and assigned physicians face a similar set of facilities, administrative resources 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 medical graduation exam scores of the team of physicians assigned to a hospital decreases the probability of giving birth to an *unhealthy* baby by 6.31%. A child is defined as unhealthy if one of these three conditions is satisfied: has a low birth weight, was an early-term infant (prematurity), or has a low Apgar score. These effects are consistent across each health measure at birth: we find a negative impact of 7.71% on low birth weight, 7.97% on prematurity, and a 7.16% decrease in the probability of low Apgar scores.[3,4] Our results are similar to the findings in shared work experience between surgeons and health care physicians (Chen, 2021) or eligibility to Medicaid for pregnant women (Currie and Gruber, 1996).[5]

To shed light on the potential channels through which physicians impact a child's outcomes, we first analyze several heterogeneous effects across different mothers' characteristics. Although the effects are slightly more pronounced among first-time mothers, teenage mothers, mothers with low education, and single women, the differences between groups are not

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

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

[4]Unfortunately, for our period of analysis, we cannot test the impact of physicians' skills on mortality because of data issues. First, we do not have the number of gestation weeks for 20% of the observations (registers) in the mortality dataset; thus, we cannot identify if those mothers were exposed to each team of physicians. Second, for 10% of the records, we do not have the ID/code of the hospital; thus, we cannot link physicians and mortality records. Finally, the mortality data set does not have information about the mothers and children. Overall, between 20% to 30% of the records are missing. We repeat the same exercise (using this data with all the limitations mentioned before) but using infant mortality as the main outcome. We find a negative effect but not statically significant at conventional levels.

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

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). To the extent that more-skilled physicians improve children's health outcomes at birth, they may help mitigate such adverse shocks in utero. The reduction in the probability of giving birth to an unhealthy child is particularly pronounced among male newborns, but it is not statistically different for male and female newborns.

Furthermore, we study heterogeneous effects between hospitals with high and low incidences of poor newborn health using ex-ante hospital-level health measures. The effects on the probability of giving birth to an unhealthy baby are larger for hospitals with a high incidence of poor newborn health, which we define as the hospitals in the top quartile of the unhealthy babies baseline incidence distribution.

We next explore a mechanism through which physicians may improve health at birth: 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.[6] We follow the standard recommendations by WHO (2016) in 2013 and define "adequate prenatal care" as having at least four visits to the doctor during pregnancy.[7] We find that more-skilled doctors, on average, do not schedule more prenatal checkups.[8]

We then test whether the more-skilled physicians target prenatal consultations toward the most vulnerable mothers, measured as those with a higher predicted likelihood of giving birth to an unhealthy baby. We use several machine learning techniques to generate two groups of predictions of the probability of giving birth to an unhealthy baby using a set of mother-hospital characteristics that are observable for the physicians at the time of prenatal care. The results show, regardless of the method we use to predict unhealthiness, how more-skilled doctors do not increase the probability of having at least the suggested number of antenatal consultations for mothers with a low predicted probability of giving birth to an unhealthy child. The doctors seem to target those prenatal checkups toward more vulnerable

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

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

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

mothers, measured as mothers with a higher predicted probability of giving birth to an unhealthy child. Consistent with prenatal care being one of the channels through which more-skilled physicians have an impact on children, we show that the effects of more-skilled physicians on birth outcomes (unhealthy, low birth weight, prematurity, and Apgar score) are particularly pronounced among mothers with an ex-ante, high predicted probability of giving birth to an unhealthy child. Altogether, these results are consistent with physicians being time constrained and unable to increase the average 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 four previous years (2009-2012) for which the doctors working at hospitals were randomly assigned (2013 and 2014). We find that the treatment generates precisely estimated zeros. Second, we show evidence on the actual randomness of the assignment by showing that our physicians' skills are not correlated with the assigned hospital and municipality's ex-ante characteristics.

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 physicians' diagnosis skills (Currie and MacLeod, 2020), physicians' teams (Chen, 2021), 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, the present paper is the first to document experimental evidence of the impact of physicians' medical skills on health outcomes.

Second, the present study contributes 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). Specifically, Abaluck et al. (2016) showed that physicians

[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.

do not target testing to the highest-risk patients, because observable risk factors receive little attention in physicians' testing decisions. In the present paper, we benefit from recent advances in machine learning techniques to show that more-skilled physicians target prenatal consultations toward mothers with the highest risk of giving birth to an unhealthy child.

Our 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 showed that access to health insurance for pregnant women lowered the incidence of low birth weight. We contribute to this literature by showing the intensive margin effects of being exposed to more-skilled doctors for those with some health coverage.

We also 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 relates to the literature on teacher value added, where the effect on students of a high-quality (effective) teacher has proven 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 having a child with an unhealthy condition by 6.3 percent. Our findings suggest that, similar to good teachers, good doctors have the potential to enhance social value through improving child outcomes at birth.

The remainder of this paper is organized as follows: In Section 2, we describe the Colombian health system and the SSO program, the setting we exploit to identify parameters of interest. Section 3 describes the rich administrative data we derive from the doctors' college exit exams and patients' outcomes at birth. In Section 4, we introduce our 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.

[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.

# 2 Institutional Background, Experimental Setting, and Physicians

## 2.1 Institutional Background

According to the Political Constitution of Colombia of 1991, access to health services is an individual basic right. The principles of the system are based on progressivity and equity in the distribution of subsidies and access to health services (Law 100, Congress of Colombia, 1993). Law 100 of 1993 introduced two types of health insurance: subsidized and contributive. The contributive regime is inclusive of formal employees (and their families) who contribute a fixed share of their employment income to the system. The subsidized regime is inclusive of poor household members who do not have formal employment.[11] By 2011, access to healthcare was close to universal; indeed, even in the poorest population, the coverage was 87%, while in rural areas it was about 88% (Páez 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[12] in Colombia was 87.7%, while the percentages of newborns with low birth weight and prematurity were 8.8% and 9.3%, respectively. Still, the system faces important challenges. In 2017, according to the United Nations Statistics Division database, the neonatal mortality rate (deaths per 1,000 live births) was 7.8 and the infant mortality rate (infant deaths per 1,000 live births) was 12.2.[13]

To become a physician in Colombia, one must study in an undergraduate medical program. Like college programs in nursing, bacteriology, and dentistry, medicine is considered a health program. Students accepted into health programs earn a BA after five to six years of education. According to Colombian law, all professionals graduating from health programs are social servants; directly after graduation, they must provide professional services in urban and rural areas 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 (Congress of Colombia, 2007), but it was only adopted in 2010 when its implementation was legislated by Resolution 1058/2010

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

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

(Ministry of Health, 2010). The main objective of the SSO is to improve the access to and quality of health services in depressed urban and rural areas (or those with difficult access to health services), as well as stimulate an adequate geographic distribution of human talent in health. The SSO also promotes spaces for the personal and professional development of those beginning their careers in the health sector.[14]

Physicians play a key role in the Colombian health system. The Ministry of Health (2013), in resolution 1441 of 2013, states that any physician in Colombia can perform low-complexity surgeries and procedures, including childbirth, C-sections, medical care to newborns, and early detection activities like antenatal consultations. An important characteristic of the Colombian health system is that physicians must carry out prenatal examinations. According to the practical guide for 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. In fact, calculations from the VSR show that physicians are responsible for all prenatal check-ups, and physicians attend 98% of all deliveries.[15]

## 2.2 The experimental setting: SSO program

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

At the end of 2012, a more organized way of running the random assignments was introduced. The first years of implementing the new SSO program proved that the directions Resolution 1058/2010 gave were not strong enough to guarantee a transparent and organized assignment of physicians. Resolution 4503/2012 (Ministry of Health, 2012a) was introduced to give clearer and more organized guidance about how the random draws should be conducted. Resolution 566/2012 (Ministry of Health, 2012b) mandated that starting in January

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

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

2013 there would be four yearly (state level) waves of SSO draws,[16] where professionals who applied to a specific state would be assigned randomly 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 to one state only and only when the number of applicants for that state remained 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 and cohort.

After the application process closes, each state runs a public, random assignment of the available spots for each profession, according to the following steps: First, an oversight board consisting of one civil servant from the state secretariat of health, and four health professionals are chosen. The civil servant then publicly announces the number of spots available and who registered for each profession. At this point, she also states the rules for the lotteries, typically by using ballots. If a health professional gets a white ballot, they are exempt from the social service and receive a certificate that allows them to work in Colombia as a professional (i.e. medical license). Otherwise, the professional gets a red ballot with the randomly assigned code of the specific hospital where they will provide their services as a professional. 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 prepares 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 one or two months after the draw and lasts for 12 months. This starting date is defined before the random assignment and therefore orthogonal to the physicians' characteristics as well. If a health professional refuses to work in the place to which they were assigned or unilaterally quits before the official end of their service, they are given a six-month sanction where they cannot work as a health professional. After that period, they must apply to the SSO program again. This sanction imposes strong costs for quitters and has proved to be a good deterrence for dropping the program.[17] The system of assigning professionals to hospitals randomly lasted for seven draws.[18] Since October 2014, a new centralized system giving more weight to professionals

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

[17] We cannot confirm whether physicians actually did work for the hospitals to which they were assigned, 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 measure may not capture all the physicians who took up the program, but it gives us a lower bound of the level of compliance with the program.

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

stating preferences and a list of prioritizations has replaced the random assignment.

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

Despite the SSO being mandatory for health graduates in different fields,[19] in this paper we focus on physicians for three reasons. First, it was the profession for which the excess demand for the state-level draws was mandatory, creating perfect conditions for lotteries. Second, in Colombia, prenatal examinations must be carried out by physicians (Gomez et al., 2013). Finally, these professionals arguably make the greatest contribution to the health of the patient (Das and Hammer, 2005), specifically to birth outcomes.

The Ministry of Health (1990, 2001) specifies that the responsibilities of a physician during their SSO period are: developing health prevention programs (such as vaccinations, family planning, antenatal controls, control of chronic diseases, buccal and visual health); providing primary care and diagnosis; assigning treatment and therapies; creating and improving medical records; making a health plan and the epidemiological profile for the local community; and performing any other duty stated in their contract. Moreover, hospitals explicitly mention attending and performing surgical procedures, including C-sections and childbirth, as part of the functions and activities of SSO physicians.[20]

# 3 Data

We use data from five main sets of administrative records. The primary dataset comes from the reports written and published by the Ministry of Health for each state-level draw implemented in January, April, July, and October 2013 and January, April, and July 2014 (Ministry of Health, 2014). From this data, we obtained individual identifications, the draw

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

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

date, the state to which physician applied, whether the student "won" the lottery or not, and notably the hospital to which each student 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 medical exams (SABER PRO). From the SABER PRO, we gleaned physicians' individual performance on two health-related fields, *health care* and *disease prevention*, plus detailed sociodemographic information about each professional.[21]

Our estimations use the scores in the two health-related exams (health care and disease prevention) as proxies of the physician's skills before the SSO program.[22] The objective of the health-related tests in the SABER PRO is to measure the skills and knowledge of medical professionals. According to ICFES, the health care module assesses whether the physician has the competence to provide care that integrates both disease prevention and proper diagnosis with medical treatment and patient rehabilitation at all levels of complexity. In addition, the module on disease prevention evaluates the physicians' competence to apply basic concepts of health promotion and disease prevention that allow the prioritization of actions according to the individuals' health conditions.

Furthermore, ICFES ranks students into one of four categories of quality. For example, the lowest level in the health care module includes students who only understand basic concepts and elements of epidemiology and public health. On the other hand, the highest level includes students who understand public health concepts (actions aimed at mitigating health problems of communities), can assess patients' health conditions, and can analyze

[21] We also get the individual performance on two other fields: reading (comprehension) and quantitative (reasoning).

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

social, cultural, and economic factors that may influence differences across patients' health. Similarly, for the disease prevention module, ICFES groups the lowest level individuals who understand basic concepts of biosafety and occupational risk. The highest level includes professionals who can analyze complex health situations in a given context and select appropriate actions following current regulations and standards in medicine. Because the SSO program is the physicians' first real work experience, and the SABER PRO is taken just before graduation, we consider their scores a good measure of the physicians' general and medical skills at the time they start their SSO service and professional career.[23]

In Colombia, as in 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 health-specific SABER PRO test scores between and within programs (and universities) for the physicians in our sample.[24] 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, because it allows us to compare the outcomes of patients who were randomly exposed to physicians with very different knowledge bases and skills.[25]

Using the scores and demographic characteristics from the SABER PRO, Guarin et al. (2021) showed that the SSO lotteries in our sample are well balanced between winners and losers. Appendix Table A.1 and Appendix Figure A.1 replicate the balancing tests in Guarin et al. (2021). Appendix Table A.1 shows individual regression between the lotteries and physicians' characteristics. In addition, Appendix 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).[26]

The third administrative dataset comes from Vital Statistics Records (VSR) collected by the Administrative Department of Statistics - DANE (Administrative Department of Statistics, 2018). The VSR records have rich information for all birth certificates filed in hospitals

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

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

[25] Similarly, Appendix Figure A.5 shows the quantitative and reading test scores for the universities the physicians in our sample attended.

[26] 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).

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 to which they were assigned. Using the birth date and number of gestation weeks from VSR, we are able to identify children born between 2013 and 2016 who were exposed to each team of physicians. We also use the VSR data from 2009 to 2012 to create mother and hospital-level controls to provide evidence of the covariate balance at the hospital level and to run falsification tests (Administrative Department of Statistics, 2018).

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

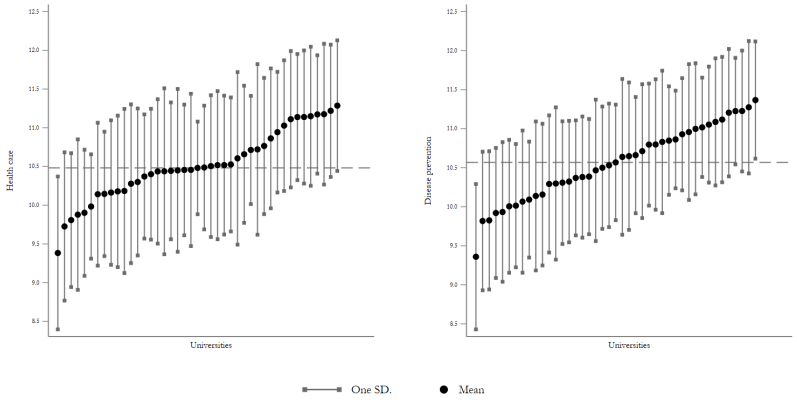
Finally, we collect information from the National Registry of Human Resources in Health, known as RETHUS. The Ministry of Health designed RETHUS through Law 1164 of 2007 (Congress of Colombia, 2007). RETHUS registers all individuals authorized to practice a profession or occupation in health. This data contains detailed information on the date of degrees, the date on which the medical license was granted, and postgraduate degrees. We also collected additional data at the 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, which are the ones that were done at random. We constraint our main sample to municipalities located outside of the main metropolitan areas for three reasons. First, the program's objective is to provide professional services in mostly rural areas with difficult access to health services (Resolution 1058/2010, Ministry of Health, 2010). Between 2013 and 2014, 77.3% of the available positions were located in small cities beyond the main 23 Colombian metropolitan areas. Second, Colombia classifies hospitals by three levels of complexity. Local municipalities usually manage level 1 hospitals. Hospitals at this level are institutions with low complexity technology, simple and easy to use in outpatient, hospitalization, emergency, and support services for diagnosis and treatment of minor health problems. Moreover, care is provided primarily by health care professionals. All hospitals in our main sample are level 1.[27] Finally, mothers located in metropolitan

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

Figure 1: Heterogeneity in SABER PRO scores in medicine programs




Notes: Figure 1 reports the health care and prevention diseases test scores for the universities that the physicians in our sample attended. Data accounts for 44 different universities. The figure shows the mean score for each university/program and an interval of one standard deviation. The dashed horizontal line represents the overall 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 and almost a one standard deviation difference between the averages of the worst and the median program and the averages of the median program and the best program.

areas have access to a large supply of hospitals at all levels, where they can easily substitute among hospitals, while mothers in small cities outside of the metropolitan areas usually only have access to one level 1 hospital. We thus expect assigned physicians to play a less pivotal role in metropolitan areas.

The municipalities included in our sample cover around 58% of the Colombian population. Finally, 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.

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

Table 1 provides the basic descriptive statistics for the main health outcomes used from the VSR. It also shows how our sample changes as we add the restrictions used in our

level of technology and care by specialized and subspecialized health professionals at all levels of care.

main estimations. Columns 1 and 2 show the mean and standard deviation, respectively, for newborns in hospitals where at least one SSO physician was assigned (SSO sample); columns 3 and 4 show the same statistics when we constrain the sample to the municipalities outside of the main metropolitan areas (i.e., rural areas). The last two columns (3 and 4) correspond to our final main sample. In our main sample, 4.26% of births were low birth weight (LBW), 4.09% were early-term infants (prematurity), and 3.74% of births had an Apgar score below 7. Our main outcome, *unhealthy*, takes the value of 1 if at least one of the previous outcomes (LBW, prematurity, or Apgar) is one. In our sample, 9.5% of the births experienced at least one of these three medical complications. Moreover, 16.28% of the mothers experienced insufficient prenatal care which is an indicator variable that takes the value of 1 if the mother received less than 4 prenatal visits. Finally, teenage pregnancy is 28.46% of total births in the main sample and the share of female newborns is 48.85%.

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

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

Notes: Table 1 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; insufficient prenatal visits is the proportion of mothers who had less than four visits; and teenage mother is the proportion of mothers aged 19 years old or less.

### 3.1.1 Municipalities

As aforementioned, we keep those municipalities in rural areas—outside of the main 23 Colombian metropolitan areas—where we expect fewer physicians per municipality. There are 600 municipalities included in our sample (see Appendix Figure A.3). The median number of people living in each municipality is 14,049 (the mean is 22,042). The average share of people living with unsatisfied basic needs (UBN) is almost 50%, with municipalities

where the whole population live under UBN.[28] This reassures that SSO physicians provide their services for one year in hospitals located in underserved areas.

We obtain the number of physicians per municipality from RETHUS.[29] From the 600 municipalities included in our sample, only 10 have more than two hospitals per municipality. The median number of physicians per hospital is 3, and around 94% of the hospitals have less than 20 physicians per hospital.[30]

Moreover, most of the deliveries are attended by general practitioners and SSO physicians. Around 90% (527 out of 588) of the municipalities with one hospital in our sample (590 out of 600) do not have an obstetrician/gynecologist working in their hospitals.[31] This is reassuring as we expect assigned SSO physicians to play a critical role in their hospitals.

### 3.1.2 SSO Physicians

The analysis includes information for 2,126 physicians who won the lottery for the seven draws implemented between 2013 and 2014. Table A.2 presents the baseline summary statistics of the physicians considered in our sample. Nearly 56% (55.8%) of the physicians are females; 0.29% live in a neighborhood classified with a socioeconomic stratum 1 or 2, whereas 36% of them live in a socioeconomic stratum 3.[32] The average household of the physicians consists of 4 people. 64.4% (63.4%) of fathers (mothers) of the SSO physician have a degree of tertiary education. Almost 45% (44.9%) of these households have a monthly income of less than three monthly minimum wages (22.9% earn less than two). Finally, the average score in the Health care score for the physicians considered in our sample is 10.4, with a maximum of 13.9, and a standard deviation of 1, and the average score in the Disease prevention for the physicians considered in our sample is 10.4, with a maximum of 13.4 and a standard deviation of 1.

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

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

[30] Figure A.4 shows the distribution of physicians per municipality for the sample of 590 municipalities with one hospital per municipality.

[31] We do not have the data on the type of specialist doctors for 2 out of the 600 municipalities.

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

# 4 Empirical Strategy

Our empirical setting focuses on a health production function that relates health outcomes at birth to physicians' medical skills. In our setting, multiple teams were assigned randomly 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 (i.e., unhealthy, low birth weight, prematurity, Apgar), using the following linear specification:

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

where $Y_{h,j,i,t}$ is the outcome of child $i$ born in hospital $h$ and exposed to a physician team $j$ at period $t$. $Z_{h,j(i,t)}$ is a score that measures the overall medical skills of the physician team $j$ that was assigned randomly to serve in hospital $h$ and whose service period intersects with child $i$'s gestation at period $t$.[33] Finally, $\gamma_d$ are draw-by-state fixed effects. The key identifying assumption behind our specification is that conditional on the draw-by-state fixed effects, $\gamma_d$, the allocation of physicians to hospital $h$ is independent of potential outcomes, $Y_{h,j,i,t}$. Thus, controlling for draw-by-state fixed effects is crucial to our identification strategy; otherwise, variation in physician quality could reflect other regional differences in the assignment of physicians to hospitals.[34] Finally, standard errors are clustered at the hospital level.

The coefficient of interest is $\beta$. Under the assumption that teams of doctors within each draw-state were assigned randomly to hospitals, $\beta$ (estimated by OLS) cleanly identifies the effect of a more-skilled team of physicians on children potentially exposed to their service in

[33] As aforementioned, we explore two different measures as proxies of the physicians' skills: the average score and the first principal component of the scores of the two health-related exams. The results are robust to this decision.

[34] We check the robustness of our estimates to including a vector of ex-ante hospital and team characteristics, $X_{h,j(i,t)}$, such as, number of inhabitants in the municipality, number of hospitals per municipality, area, 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. We also include a vector of sociodemographic information of mother-child $i$, $W_i$, including 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, and marital status. The results, as expected given the random assignment, show that the estimated effects are robust to the inclusion/exclusion of controls. Because we use data for three years of the infants' vital statistics, we also include year fixed effects to control for changes over time.

the assigned hospital $h$. To make the interpretation of the estimated coefficient $\beta$ straightforward, $Z_{h,t,i}$ is expressed in standard deviations of the skill measure. Therefore, the final result is interpreted as the percentage points change in the outcome variable associated with 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.

We focus on unhealthy as our principal measure of health at birth, which captures the three main outcomes on which the literature has focused on: low birth weight, prematurity, and the Apgar score. Low birth weight is defined as being born with a birth weight below 2,500 grams and has been one of the principal measures of health at birth studied in the literature (Currie, 2011). Prematurity is defined as being born before the 37th gestational week. Prematurity is highly correlated with low birth weight and mortality (Almond et al., 2005; Gonzalez and Gilleskie, 2017). Children born prematurely are at greater risk of suffering a variety of health problems, some of which can ultimately cause death.[35] For Apgar, we use an indicator of whether the newborn had a score below 7 in Apgar 1, as the threshold of 7 is commonly used in the literature (Ehrenstein, 2009).[36] Almond et al. (2005) argued that using the Apgar score to evaluate birth outcomes has the same practical advantages as birth weight: (i) it is relatively easy to collect; (ii) it is already available in birth records' data; and (iii) it is a measure that does not depend on a rare event (such as mortality).[37]

We focus on the average score of the two health-related health exams. Nonetheless, we provide robustness results using the first principal component and each score individually.[38]

[35]Complications include immunological, respiratory, central nervous system, gastrointestinal, hearing, and vision problems, as well as cognitive, motor, social-emotional, behavioral, and long-term growth problems (Butler 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.

[36]Apgar has been used in the literature as a measure of newborn health status; for example, see Almond et al. (2010) and Lin (2009). Apgar is a measurement of the health of newborns based on breathing, heart rate, color, reflexes, and muscle tone (Moore 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).

[37]Similarly, Ma and Finch (2010) recommended always including the Apgar score, because it appears to be the strongest predictor of neonatal mortality, regardless of birth weight.

[38]We repeat our main empirical analysis using 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

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.[39] We focus on municipalities outside of the main metropolitan areas for the entirety of the analysis.

Finally, to evaluate the internal validity of our identification strategy, we implement the following falsification tests: We assign a "placebo treatment" to the newborns who show up in the VSR of the four years before the program (2009, 2010, 2011, and 2012) instead of years 2013, 2014, 2015, and 2016 used in our main estimation sample. We use the same draw date, proposed start date, and hospital to which each of the physicians was assigned randomly but four years before the actual date. We then run equation (1) under the same conditions used for the main sample.

# 5 Results

This section describes the causal effects of physicians' medical skills on birth outcomes. We first test whether hospitals' birth outcomes and additional covariates measured in years 2010, 2011 and 2012 from VSR are correlated with the skills of the physicians randomly assigned in 2013 and 2014 and who provided medical care to individuals who were born between 2013 to 2016 (four years later). Our results show that there is no correlation between different health outcomes and our proxy for physicians' skills. Second, we find that physicians' skills have a negative and significant effect on our main measure, unhealthy, as well as on low birth weight, prematurity, and Apgar. Third, we provide robustness checks to our main results by using a standardized principal component and each individual score as a proxy for physicians' skill and including a large set of controls. We also rescale our measure of physicians' skills, weighting the average score of physicians by the fraction of the workforce in each hospital that arrives via the lottery. Fourth, we implement a placebo test using data for the four previous years of our main sample. Finally, we estimate heterogeneous effects on mothers' and hospitals' characteristics.

content. The quantitative test measures general knowledge in mathematics, statistics, and data analysis.

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

## 5.1 Characteristics of the hospitals and physicians' skills

To test whether the main health at birth outcomes and additional covariates, measured before our main sample of the SSO program, are correlated with the quality of the physicians assigned to each hospital, we regress each hospital's characteristics three years before our main sample of the SSO program (i.e., from 2010 to 2012) on physicians' overall skills—proxied by the average of health-related college examination scores.[40] We include draw-by-state fixed effects and cluster the standard errors at the hospital level. Table 2 shows the coefficients and their standard errors from each regression. From Table 2, it follows that there is no correlation between overall skills of physicians randomly assigned during our main sample period (i.e., in 2013 and 2014) and the health outcomes, as well as the hospital characteristics measured three years before.[41]

## 5.2 Main results on health at birth

In this section, we provide our main results on health at birth outcomes. Table 3 presents the estimated coefficient $\beta$, in equation (1), using ordinary least squares. We find that our main skill measure has a negative and significant effect on unhealthy as well as each of the health outcomes (i.e., LBW, prematurity, and Apgar). The coefficient represents the percentage points effect of an increase of one standard deviation of physicians' average health score. The standard error of the coefficient is presented in parenthesis, and below we present the relative (percent) effect (i.e., we divide the main coefficient by the average of the dependent variable).

In column (1) of Table 3, we see a significant negative relationship between physicians' skills and unhealthy—a decrease in the probability of being born unhealthy of 0.6 percentage points. Our estimates suggest that an increase of one standard deviation in physicians' average score decreases the probability of being born unhealthy by 6.31%.[42] Columns (2) to

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

[41] In the Appendix, we estimate the same regression using the average of the four fields of the college examination scores as a proxy for physicians' skills. Appendix Table A.3 shows that there is no correlation between the overall skills of physicians that arrived during our main sample period (i.e., in 2013 to 2014) and the different hospital characteristics measured three years before.

[42] In the education context, the teacher value-added literature (e.g., Chetty et al., 2014; Rothstein, 2017) found that an increase in teacher quality of one standard deviation corresponded to an increase in students' test scores of 0.19 standard deviations in math and 0.14 standard deviations in reading. Our results suggest an increase in physician quality of one standard deviation corresponds to a decrease in the probability of being born unhealthy by 6.31%. We find similar effects (6.81%) when we use the average of reading and

Table 2: Covariate balance at hospital level

| Covariate | Coefficient | Standard Error |
|---|---|---|
| Unhealthy | 0.001 | 0.001 |
| Low birth weight | 0.000 | 0.001 |
| Prematurity | 0.000 | 0.007 |
| Apgar < 7 | 0.003 | 0.009 |
| Antenatal consultations < 4 (Prop.) | 0.000 | 0.003 |
| Proportion of female newborns | 0.000 | 0.001 |
| Proportion of mothers with basic education | -0.002 | 0.003 |
| Proportion of married mothers | 0.001 | 0.002 |
| Proportion of teenage mothers | 0.000 | 0.002 |
| Mean number of antenatal consultations | -0.005 | 0.022 |
| Hospitals by municipalities | 0.000 | 0.010 |
| Municipality population | 325.7 | 1,032.3 |

Notes: Table 2 reports the results of regressing each hospital's ex-ante characteristics on physicians' overall skills. Hospitals' characteristics come from the 2010-2012 DANE VSR. Low birth weight is the proportion of newborns with low birth weight (weight <2,500 grams); prematurity is the proportion of newborns who were premature (fewer than 37 weeks of gestation); Apgar is the proportion of newborns whose Apgar score is lower than 7; antenatal consultations $\leq 4$ is the proportion of mothers who had less than four 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. We interpret the non-significance of these estimates as evidence in favor of the randomness of the assignment of physicians.

(4) in Table 3 examine each measure of health at birth. The point estimate for the average of the health-specific scores is associated with a decrease in the probability of low birth weight of 0.33 percentage points (7.71%), being premature of 0.33 percentage points (7.97%) and a drop in the probability of being born with an Apgar score below 7 of 0.27 percentage points (7.16%). These results are consistent with previous literature that has found that

quantitative score as a proxy for physicians' skills (see Appendix Table A.8). Note that in our context, a one standard deviation increase is almost equivalent to the change from having a physician from the bottom-ranked program to having a physician from a median-ranked program or from having a physician from a median-ranked program to having a physician from the top-ranked program (see Figure 1).

prematurity is an important determinant of weight at birth (Almond et al., 2005).[43,44]

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

Table 3: Main estimates using all sample and average score

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

Notes: Table 3 shows our main estimates. The coefficients represent the effect of an increase of one standard deviation of the physicians' skill measure (scores). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. 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. * Significant 10%, ** significant 5%, *** significant 1%

[43]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 2009 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.

[44]We repeat the same exercise (with all the limitations mentioned before) using infant mortality as the main outcome. We find a negative effect but not statically significant at conventional levels. We also test if more skilled physicians may have affected fertility in the municipalities they were randomly assigned. Thus, we repeat the same exercise using the number of pregnancies in each municipality as the outcome. We find no evidence that more skilled physicians affected fertility. Overall, We find no evidence of selective fertility or selective child mortality.

## 5.3 Robustness Checks

We run additional specifications in which we use the standardized principal component instead of the standardized average health related scores as a proxy for physicians' skills. In addition, we show that our results are robust to the inclusion of ex-ante hospital characteristics as well as a vector of sociodemographic information of mother-child. Figure 2 compares the estimated coefficient (relative to the mean), $\beta$, in equation (1) using the average (main specification) and the principal component of health scores with and without controls. We see from Figure 2 that our estimates for unhealthy—and each of the three health measures—are similar if we use the first principal component as a proxy for skills and are robust to the set of controls included in our analysis.[45,46]

Note that the average prevalence of the outcomes considered is usually low and around 4%. One concern might be that a linear regression may not fit the data well. To alleviate this concern, we estimate equation (1) using an analogous Logit model and compute the average marginal effect associated with an increase in one standard deviation of the skill measure. Appendix Table A.6 shows that the marginal effects (signs and magnitudes) are very similar to the ones estimated using a linear regression model.

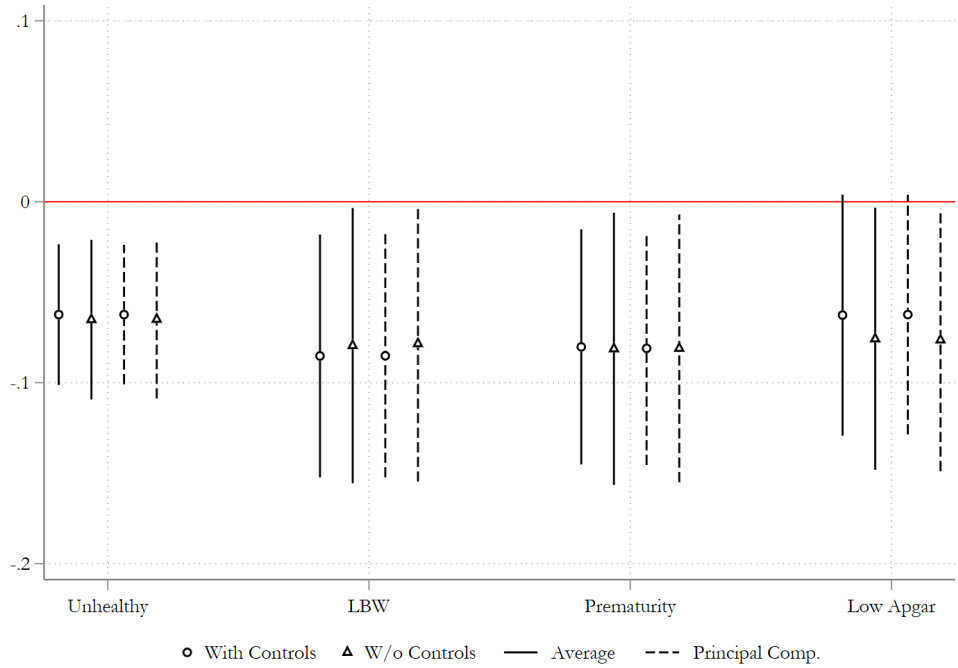
Also, while ordinary least squares allows us to compute the average effect of our skills measure, it does not tell us much about the magnitude of this effect across the distribution of skills. We rank the score into quartiles and estimate equation (1) using a set of dummies indicating the score distribution quartile to which physicians belonged. The results are presented in Appendix Table A.7. Columns (1) and (2) present the coefficients associated with the effect of belonging to the second, third, and fourth quartile of the distribution of the average of the health-related scores and the first principal component, respectively, on our main outcome, unhealthy, relative to the first quartile. Although we lack power to find statistically significant differences, we see that the point estimates are negative and monotonically decreasing with respect to the quartile. This suggests that there are potential gains associated with getting a more-skilled physician across the whole distribution of skills.

Finally, we extend our analysis by estimating our main specification using alternative

[45] Results are reported in Appendix Table A.4, where we use unhealthy, low birth weight, prematurity, and Apgar score as our dependent variables, using the standardized average health-related college examination score and standardized principal component as a proxy for physicians' skills, with and without controls.

[46] We standardized, centered, and aggregate the three main health outcomes (LBW, prematurity, and Apgar score) using the inverse covariance index suggested by Anderson (2008) and repeat our main empirical analysis using the index as dependent variable. Appendix Table A.5 presents the results using the covariance index and our main outcome, unhealthy (standardized), as dependent variables. Note that the adjusted standardized coefficients (in standard deviations) are very similar for both specifications.

Figure 2: Main estimates using all sample




Notes: Figure 2 presents the coefficients for the relative effect of an increase of one standard deviation of the physicians' 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7, and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status, number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 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 skills. 95% confidence intervals.

measures of skills. We use the average of the four areas tested in the SABER PRO (health management, public health, reading, quantitative), as well as each individual score as proxies of the physicians' skills before the SSO program. We regress unhealthy on the different proxies for physicians' skills, and show in Table A.8 that the scores have a negative effect on

unhealthy and are not statistically different from each other.[47]

### 5.3.1 Weighted score

In our main regression (equation 1), the coefficient on the average quality may underestimate the effect of doctor quality if the randomly assigned doctors are not the entire workforce of the hospitals. Moreover, the underestimation depends on the fraction of the workforce in each hospital that arrives through the lottery.

To quantitatively explore this idea, we rescale our measure of physicians' skills, weighting the average score of physicians assigned to a hospital by the fraction of the workforce in each hospital that arrives via the lottery. Thus, if a hospital has a small share of doctors that are replaced by the lottery, the difference in mean outcomes caused by the quality difference of the SSO doctors will be much smaller. We excluded municipalities with more than two hospitals per municipality (10 municipalities), because we cannot identify the hospitals where non-SSO doctors work.[48]

Table 4 reports the results of the estimation using the weighted score.[49] The table shows that the results are very similar when the weighted score is used to proxy physicians' skills. These results are consistent with the fact that the median number of doctors per hospital is three, and around 95% of the hospitals have less than 20 doctors per hospital.

[47] In Appendix Table A.9, we interact the average score with the university's (program) average score to test if top universities drive the estimated effect. We do not find evidence that top-ranked universities drive the effects presented before.

[48] Appendix Table A.10 shows the results excluding from our sample the ten municipalities with more than two hospitals per municipality, using our main unweighted proxy for physicians' skills. Reassuringly, this restriction delivers results that are very similar to our baseline results (Appendix Table A.4).

[49] Appendix Table A.11 reports the results using unhealthy, low birth weight, prematurity, and Apgar score as our dependent variables with and without controls.

Table 4: Main results with weighted score

| | **Unhealthy** | **LBW** | **Prematurity** | **Apgar < 7** |
|---|---|---|---|---|
| | | Average Health Scores | | |
| | (1) | (2) | (3) | (4) |
| Coefficient | -0.0064*** | -0.0034* | -0.0037** | -0.0029** |
| Stand. Err. | (0.0022) | (0.0018) | (0.0017) | (0.0013) |
| Adjusted Coeff. | -6.66% | -7.78% | -8.80% | -7.63% |
| Average Dependent Variable | 0.096 | 0.043 | 0.042 | 0.037 |
| Number of Observations | | 237,082 | | |

Notes: Table 4 shows our main estimates weighting the average score of physicians assigned to a hospital by the fraction of the workforce in each hospital that arrives via the lottery. We excluded municipalities with more than two hospitals per municipality (10 municipalities). The coefficients represent the effect of an increase of one standard deviation of the physician skill measure (scores). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state fixed effects. 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.
* Significant 10%, ** significant 5%, *** significant 1%

## 5.4 Placebo Tests

To evaluate our identification strategy's validity, we implement a placebo test, using VSR records for children born in 2009, 2010, 2011, and 2012. Recall that for our main results, we use data from the physicians randomly assigned in 2013 and 2014 and who provided medical care to individuals who were born between 2013 to 2016. We move the physician's arrival time four years back and run placebo tests similar to our main specification but using data for the four previous years (2009-2012). We then estimate equation (1) using the same outcomes and set of fixed effects used in Table 3.

Because physicians in our sample did not treat children born in 2009, 2010, 2011, and 2012, we would expect a null effect. Table 5 shows that the point estimates are precisely estimated zeros for our main outcome, unhealthy, and for each of the other health outcomes (LBW, prematurity, Apgar).[50] Our results are robust to the use of the first principal com-

[50]In Appendix Table A.12, we repeat the same exercise and present the results for windows of 3.5, 3, 2.5

ponent as a proxy for skill, as well as to the inclusion of a set of controls such as ex-ante hospital and team characteristics as well as a vector of sociodemographic information of mother-child (Appendix Figure A.6 and Table A.13). Finally, 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.14 shows that the marginal effects (signs and magnitudes) are null to the ones estimated using a linear regression model.

Table 5: Placebo test

| | **Unhealthy** | **LBW** | **Prematurity** | **Apgar < 7** |
|---|---|---|---|---|
| | **Health Average Score** | | | |
| Coefficient | -0.0009 | -0.0008 | -0.0020 | 0.0004 |
| Stand. Err. | (0.0022) | (0.0011) | (0.0016) | (0.0013) |
| Adjusted Coeff. | -0.79% | -1.76% | -3.76% | 0.78% |
| Average Dependent Variable | 0.119 | 0.046 | 0.052 | 0.047 |
| Number of Observations | 262,089 | | | |

Notes: Table 5 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 physicians' skill measure (average score). Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state fixed effects. 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.
* Significant 10%, ** significant 5%, *** significant 1%

## 5.5 Physicians' impacts across subgroups

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

and 2 years before the start of the SSO program.

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

We find that the effect of physicians' skills on our main outcome, unhealthy, is slightly more pronounced among first-time mothers and teenage mothers (see Table 6), but we do not find statistically significant differences on the effects across mothers' characteristics. Furthermore, we do not find statistically significant differences across mothers with high and low education, as well as married and single mothers (Appendix Table A.15). Finally, 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).[51] It is possible that skilled physicians play an important role in mitigating negative shocks on more vulnerable fetuses. Although we find that the reduction in unhealthy was particularly pronounced among male newborns, we do not find any statistical difference between males and females (see Appendix Table A.15).

### 5.5.1 Hospitals' characteristics

Finally, we look at heterogeneity across hospitals' characteristics. We divide the sample between hospitals below (low incidence) and above (high incidence) the 75th percentile of our main outcome—unhealthy—distribution using data from the SSO program for the three years before our sample period (2010-2012). In Table 6, columns 1 and 2, we test the effects associated with physicians assigned to hospitals with a low or high incidence of unhealthiness for these three years.

We find a (weak) significant difference between the effect of physicians' skills on unhealthy in hospitals with a high and low incidence of poor health. The effect is strongly negative and significant in hospitals with a high incidence of poor health (Currie, 2011). The point estimate for physicians in hospitals with high (low) incidence is -0.73 (-0.41) percentage points. An increase of one standard deviation in physicians' average score decreases the probability of a child being born unhealthy by 6.08% (4.32%) in a hospital with high (low)

[51] 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).

Table 6: Heterogeneity of the effects across mothers' characteristics

| | Unhealthy | | | | | |
|---|---|---|---|---|---|---|
| | Hospital | | Mother | | | |
| | Higher incidence of Unhealthy | Lower incidence of Unhealthy | First-time | Non-first-time | Teenage mothers | Non-teenage mothers |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | **Average score** | | | | | |
| Coefficient | -0.0073** | -0.0041* | -0.0070*** | -0.0055*** | -0.0070*** | -0.0058*** |
| Stand. Err. | (0.0031) | (0.0022) | (0.0025) | (0.0018) | (0.0024) | (0.0019) |
| Adjusted Coeff. | -6.08% | -4.32% | -6.40% | -5.78% | -6.25% | -6.12% |
| Average Dependent Variable | 0.120 | 0.095 | 0.109 | 0.095 | 0.113 | 0.095 |
| Number of Observations | 101,556 | | | | | |

Notes: Table 6 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 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 fixed effects. 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.
* Significant 10%, ** significant 5%, *** significant 1%

incidence of unhealthy. Reassuringly, this evidence suggests that physicians play a more important role in hospitals with a history of poor health outcomes.[52]

# 6 Potential Mechanism

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, such as the quality of medical advice doctors provide, are unobservable (Das et al., 2008; Leonard and Masatu, 2007), whereas others, such as the number of prenatal consultations, are observable. In this section, we study prenatal consultations as a potential mechanism for observed differences between skilled and unskilled

[52]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) found significant health benefits for older patients who were brought to higher-cost hospitals, Card et al. (2019) found that, during the first year of life, newborns who were delivered by C-section were more likely to visit the emergency department, less likely to be readmitted to hospital, and had 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.

physicians.

## 6.1 Prenatal consultations

We first explore whether more-skilled physicians increase the number of prenatal consultations, serving as a mechanism to improve the quality of care and health outcomes. Although most of the body of evidence from both economics and medical research shows an important association between prenatal care and both birth weight and prematurity, there are some disagreements (Alexander and Korenbrot, 1995; Amarante et al., 2016; Carrillo and Feres, 2019; Conway and Deb, 2005; Currie and Grogger, 2002; Grossman and Joyce, 1990; Kramer, 1987; McCormick and Siegel, 2001).[53]

According to WHO (2016) and the Colombian government (Gomez et al., 2013), prenatal care improves the health status of both mother and newborn. In Colombia, the Ministry of Health requires physicians to carry out prenatal monitoring (Gomez et al., 2013). We follow the standard recommended by WHO (2016) for our period of analysis and measure "adequate prenatal care" contact as having at least four visits to the doctor during pregnancy. We do not find evidence that more-skilled doctors reduce the probability that mothers are scheduled for less than four prenatal checkups (see Appendix Table A.16).

We expect that physicians enrolled in the SSO program and assigned to rural areas (outside the metropolitan areas) would be time constrained, as usually they are the only physicians available in those areas.[54] Anecdotal evidence supports this notion, as described in various reports from Colombian medical associations in which physicians refer to the SSO year as an experience during which they had an overwhelming workload and long working hours.[55] In this setting, in which 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 most likely to give birth to an unhealthy baby.

We assume that the probability of an unhealthy baby can be thought of as a prediction

[53] 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.

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

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

problem and take advantage of recent advances in machine learning techniques.[56] We use these techniques to generate two groups of predictions about the mothers' probability of giving birth to an unhealthy baby 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 and logistic regression models.[57]

The sample is clustered into training and testing subsets of randomly selected hospitals using K-means algorithm. We repeat this procedure—splitting the main sample using K-means—5,000 times. We run Logit and random forest models on the training sets and use the models to predict the probability of giving birth to an unhealthy child on each testing subset.[58] We then divide the test sample into two groups: *low* and *high* predicted probability, defined as mothers with a probability of giving birth to an unhealthy child below and above the 75th percentile, respectively, for each of the two model predictions.[59]

We estimate equation (1) using a dummy that is equal to 1 if the number of prenatal consultations is less than four—as our main outcome—in each of the previously defined groups (i.e., low and high predicted probability of an unhealthy child). Table 7 presents the average coefficient and the standard error for the 5,000 repetitions.[60] Columns (1) and (2) present the results for the sample of mothers with a low predicted probability, and columns (3) and (4) for the sample of mothers with a high predicted probability of giving birth to an unhealthy child. We include the results both with and without controls.

Table 7 shows that regardless of the method we use, more-skilled doctors do not seem to increase the recommended number of antenatal consultations for mothers with a low predicted probability of giving birth to an unhealthy child. Instead, they target prenatal checkups toward the more vulnerable mothers, measured as mothers with a high predicted

[56] 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.

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

[58] We follow Chernozhukov et al. (2018) and re-scale the outcomes and covariates to be between 0 and 1 before training.

[59] Liberman et al. (2018) and Liberman et al. (2021) followed a similar strategy when they studied the effects of information deletion and usury rates on consumer credit markets.

[60] Figure A.7 shows that the distribution of the estimated coefficients for all the 5,000 repetitions.

probability of giving birth to an unhealthy baby. Consistent with our suggested mechanism of physicians being able to target care toward the more vulnerable mothers, we find stronger effects of our measure of skills when we focus on mothers with a higher predicted probability compared to those with lower predicted probability. While the point estimate for the effect of physicians' skills on unhealthy in the lower predicted probability sample is between -0.08 and 0.09 percentage points depending on the prediction used to divide the data, the point estimate for the higher predicted probability group is between -1.3 and -0.91 percentage points. These estimates suggest that an increase of one standard deviation in physicians' average score decreases the probability that mothers are scheduled for less than four prenatal checkups between 5.59% and 7.98% for mothers with high predicted probability of giving birth to an unhealthy child.

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

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

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

Notes: Table 7 reports the differential effects of physicians' skills measure on antenatal consultations by mother's predicted probability of giving birth to an unhealthy child. To predict the probability of an unhealthy child, we divided our data into training and testing subsets of randomly selected hospitals using K-mean algorithm. On the training sets, we run Logit and random forest models, and use the estimations to predict the probability of giving birth to an unhealthy child on each testing subset. Using the prediction on the testing sample, we divide each subset into high and low predicted probability of giving birth to an unhealthy child, defined as mothers with a probability of an unhealthy child below and above the median, respectively. The coefficients presented represent the effect of 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 insufficient (less than four) antenatal consultations. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. All regressions control for draw state. Numbers in parentheses are clustered standard errors. The results show that regardless of the method we use to predict unhealthy babies, more-skilled doctors do not seem to increase the recommended number of antenatal consultations for mothers with a low predicted probability of giving birth to an unhealthy child. Instead, they target those prenatal checkups toward the more vulnerable mothers.
* Significant 10%, ** significant 5%, *** significant 1%

### 6.1.1 Effect on unhealthy

We next show —consistent with the idea of better physicians being better at targeting care to the most vulnerable mothers—if more skilled physicians reduce the probability of giving birth to an unhealthy child, particularly among the most vulnerable mothers. Table 8 shows that more skilled doctors seem to improve health at birth of children for all mothers (i.e., with a low and high predicted probability of unhealthy babies). However, the effect is more pronounced, regardless of the method we use to split the sample, for mothers with a (ex-ante) high predicted probability of unhealthy babies. In particular, for the more vulnerable mothers, an increase of one standard deviation in physicians' average college examination score decreases the probability of an unhealthy newborn around 9%, while for mothers with (ex-ante) low predicted probability of an unhealthy newborn, the effects are smaller in

magnitude, close to 5%.[61]

Table 8: Main outcomes by predicted unhealthiness

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

Notes: Table 8 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 Table 7. 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. 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.
* Significant 10%, ** significant 5%, *** significant 1%

# 7 Conclusions

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

In Colombia, medical school graduates must spend the first year of their careers working in the national Mandatory Social Service program (SSO). The SSO program randomly assigns physicians to their first job, providing a test for the effects of being treated by a more-skilled

[61] Figure A.8 presents the distribution of the estimated coefficients for the 5,000 repetitions for the four outcomes studied.

physician. In this paper, we combine administrative records to match physicians in the SSO program, hospitals, vital statistics records, characteristics of the physicians, and mandatory health-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 giving birth to an unhealthy child. We estimate that an increase in one standard deviation in the physicians' academic health test score reduces the probability of giving birth to an unhealthy child by 6.31%. Although unhealthy is our main measure of health at birth, the results are robust to other measures such as low birth weight, prematurity and Apgar score.

Furthermore, we explore whether 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 for more prenatal checkups. Nonetheless, we provide evidence that these physicians are targeting their prenatal consultations toward the most vulnerable mothers, measured as those with the predicted likelihood of giving birth to an unhealthy baby.

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

# References


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

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

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

Administrative Department of Statistics (2005). National census.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Black, S. E., P. J. Devereux, and K. G. Salvanes (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., R. E. Behrman, et al. (2007). *Preterm birth: causes, consequences, and prevention*. National Academies Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gomez, P., I. Arevalo, et al. (2013). Guías de práctica clínica para la prevención, detección

temprana y tratamiento de las complicaciones del embarazo, parto y puerperio. *Ministerio de Salud y protección social Colombia 84*, 74–82.

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

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

Guarin, A., C. Posso, E. Saravia, and J. Tamayo (2021). Healing the gender gap: The impacts of randomized first-job on female physicians.

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

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

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

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

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

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

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

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

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

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

Liberman, A., C. Medina, C. Neilson, and C. Posso (2021). Lender market power and the

bright side of interest rate caps: Evidence from colombia. Technical report, Unpublished manuscript.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Veddovi, M., D. T. Kenny, F. Gibson, J. Bowen, and D. Starte (2001). The relationship be-

tween depressive symptoms following premature birth, mothers' coping style, and knowledge of infant development. *Journal of Reproductive and Infant Psychology 19*(4), 313–323.

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

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

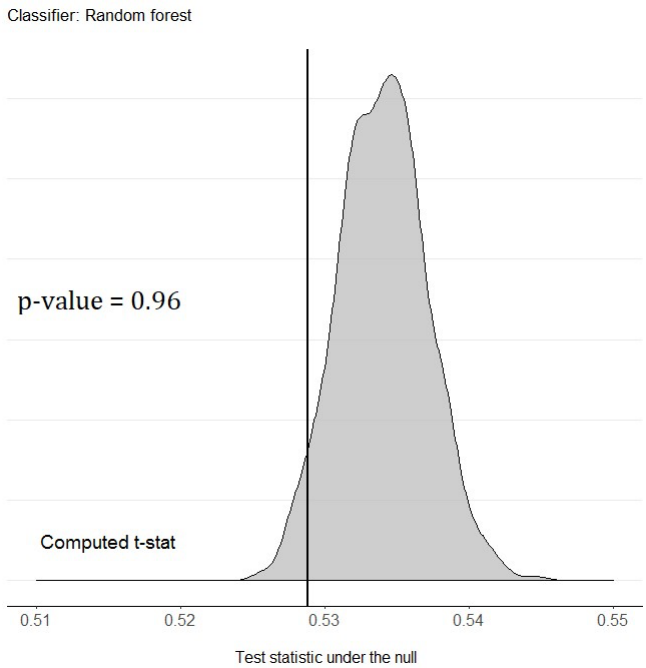
Woodcock, S. D. (2008). Wage differentials in the presence of unobserved worker, firm, and match heterogeneity. *Labour Economics 15*(4), 771–793.

# Online Appendix

Not for Publication

# A Appendix

Figure A.1: Balancing test using the classification permutation test (Gagnon-Bartsch et al., 2019)



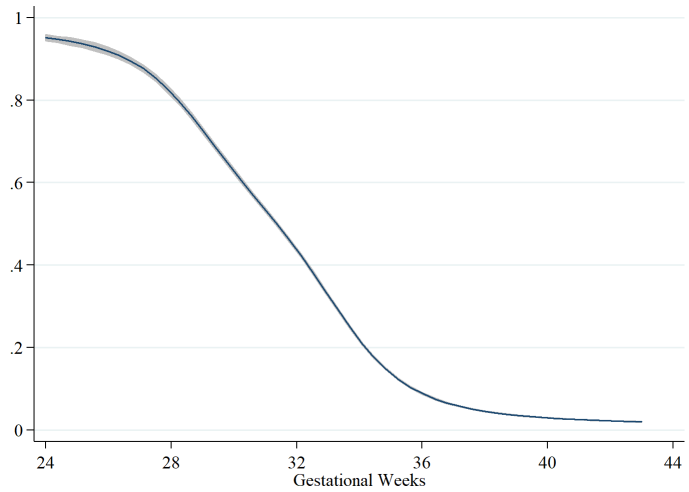

Notes: Figure A.1 shows the results for the Classification Permutation Test: A Machine Learning Nonparametric Test for Equality of Multivariate Distributions (Johann Gagnon-Bartsch and Yotam Shem-Tov, 2018, *Annals of Applied Statistics*). The procedure includes 1,000 repetitions. We also perform a reverse regression test (F(19, 160) = 0.88, p-value = 0.6128). These results 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 SABER PRO 4 | 10.539 | 0.833 | 0.007 | 0.028 |

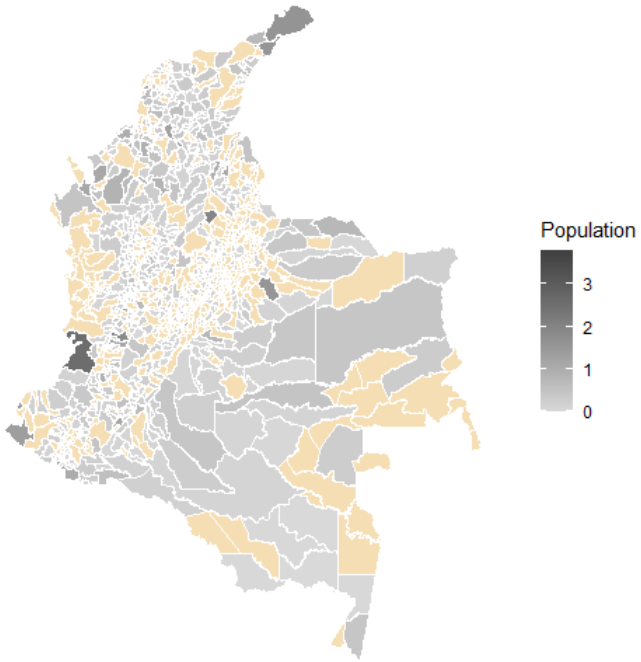
Notes: Table A.1 reports lottery losers' means and estimated effects of winning the SSO, based on a sample of 3,559 observations with a 3,519-degree of freedom, testing a total of 20 hypotheses. Standard errors are clustered, given the by draw and state design of the randomization. Controls for draw-by-state 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.

Figure A.2: Probability of low birth weight vs. gestational weeks, 2009-2012



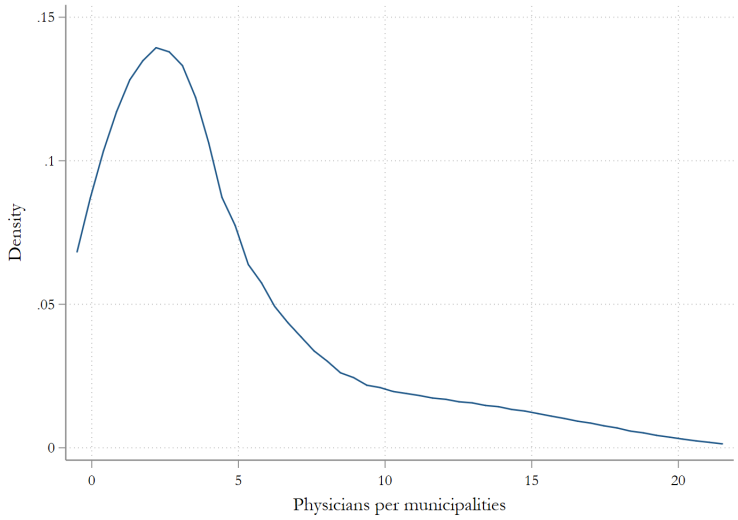

Notes: Figure A.2 presents the local polynomial regression fit of the probability of having low birth weight over the number of gestational weeks using all birth records for Colombia from 2009 to 2012.

Figure A.3: Population (per 100,000) for municipalities included in our main sample




Notes: Figure A.3 presents the map of the population per 100,000 people for the municipalities included in our main sample in 2005. The municipalities in orange are not included in our sample or do not have SSO.

Figure A.4: Distribution of physicians per municipalities




Notes: Figure A.4 shows the distribution of physicians per municipality for the sample of 590 municipalities with only one hospital. The data spans from January 2012 to December 2012.

Figure A.5: Heterogeneity in quantitative and reading SABER PRO scores

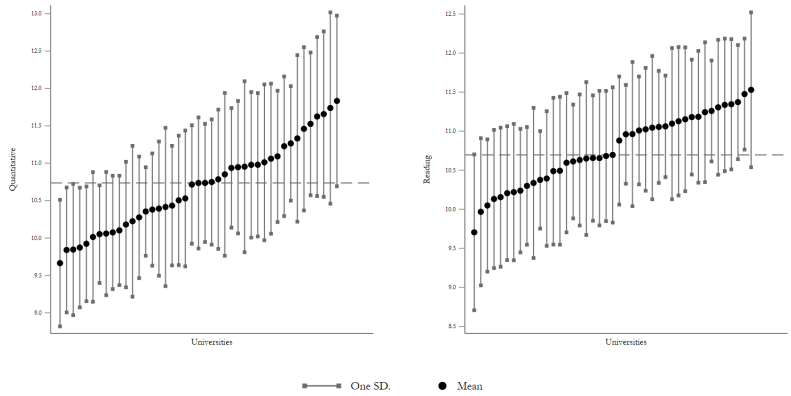



Figure A.5 reports the quantitative and reading test scores for the universities that the physicians in our sample attended. Data accounts for 44 different universities. The figure shows the mean score for each university/program and an interval of one standard deviation to each side of the average. The dashed horizontal line represents the overall percentile 50. The figure shows substantial heterogeneity both within and between programs. For all the fields reported, there is a difference of almost two standard deviations between the averages of the best and the worst programs.

Table A.2: Summary statistics - physicians in the main sample

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

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

Table A.3: Covariate balance at hospital level using all the areas tested in the SABER PRO

| **Covariate** | **Coefficient** | **Standar Error** |
|---|---|---|
| Unhealthy | 0.001 | 0.001 |
| Low birth weight | 0.000 | 0.001 |
| Prematurity | 0.000 | 0.007 |
| Apgar $< 7$ | 0.003 | 0.009 |
| Antenatal consultations $< 4$ (Prop.) | 0.000 | 0.003 |
| Proportion of female newborns | 0.000 | 0.001 |
| Proportion of mothers with basic education | -0.002 | 0.003 |
| Proportion of married mothers | 0.001 | 0.002 |
| Proportion of teenage mothers | 0.000 | 0.002 |
| Mean number of antenatal consultations | -0.005 | 0.022 |
| Hospitals by municipalities | 0.000 | 0.010 |
| Municipality population | 325.7 | 1,032.3 |

Notes: Table A.3 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 $\leq 4$ is the proportion of mothers who had less than four 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. We interpret the non-significance of these estimates as evidence in favor of the randomness of the assignment of physicians.

Table A.4: Main estimates without and with controls

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

Notes: Table A.4 presents the main results with and without controls. 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status; number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to the inclusion/exclusion of controls and the way we measure of skills. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

Table A.5: Main results using covariance index (Anderson, 2008)

| | **Unhealthy Cov index** | | **Unhealthy standarized** | |
|---|---|---|---|---|
| | **Score average** | **PCA score** | **Score average** | **PCA score** |
| | | **Panel A. Without controls** | | |
| Coefficient | -0.0160*** | -0.0160*** | -0.0211*** | -0.0210*** |
| Stand. Err. | (0.0055) | (0.0055) | (0.0072) | (0.0072) |
| Adjusted Coeff. Sd. | -2.35% | -2.35% | -2.11% | -2.10% |
| | | **Panel B. With controls** | | |
| Coefficient | -0.0153*** | -0.0153*** | -0.0202*** | -0.0202*** |
| Standard Error | (0.0050) | (0.0050) | (0.0064) | (0.0064) |
| Adjusted Coeff. Sd. | -2.24% | -2.24% | -2.02% | -2.02% |
| Number of Observations | | 256,805 | | |

Notes: Table A.5 presents the main results using covariance index. 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status; number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to using the covariance index as an outcome instead of unhealthy. The results are also robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

Table A.6: Main estimates using a Logit model

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

Notes: Table A.6 presents the main results using a Logit model. 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status; number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to using an analogous Logit model and compute the average marginal effect associated with an increase in one standard deviation of the skill measure. The results are also robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

Table A.7: Main estimates linearity

| | | Unhealthy | |
|---|---|---|---|
| | | **Health average score** (1) | **Health PCA score** (2) |
| **Quartile 2** | Coefficient | -0.0066* | -0.0070* |
| | Stand. Err. | (0.0036) | (0.0040) |
| | Adjusted Coeff. | -6.95% | -7.36% |
| **Quartile 3** | Coefficient | -0.0082** | -0.0079** |
| | Stand. Err. | (0.0039) | (0.0040) |
| | Adjusted Coeff. | -8.64% | -8.33% |
| **Quartile 4** | Coefficient | -0.0133*** | -0.0134*** |
| | Stand. Err. | (0.0036) | (0.0035) |
| | Adjusted Coeff. | -13.98% | -14.09% |

Notes: Table A.7 presents estimates using the quartiles of the skills distribution. The coefficients represent the effect of being assigned a physician of the quartiles 2, 3, or 4 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7, and zero otherwise. All regressions control for draw state fixed effects. 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.
* Significant 10%, ** significant 5%, *** significant 1%

Table A.8: Main estimates using all the areas tested in the SABER PRO

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

Notes: Table A.8 presents the main results using all areas tested in the SABER PRO. The coefficients represent the effect of an increase of one standard deviation of the physicians' skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status; number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to using the average of the four areas tested in the SABER PRO (health management, public health, reading, quantitative) as well as each individual (except for reading) score as proxies of the physician's skills before the SSO program. The results are also robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

Table A.9: Interaction between cohort scores and program scores

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

Notes: Table A.9 presents the main results using the interaction between cohort and program scores. The coefficients represent the effect of an increase of one standard deviation of the physicians' skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status; number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. These results show the effects presented in Table 3 are driven by top-ranked universities. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

Table A.10: Main results using municipalities with one hospital

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

* Significant 10%, ** significant 5%, *** significant 1% Notes: Table A.10 presents the main results using municipalities with only one hospital. The coefficients represent the effect of an increase of one standard deviation of the physicians' skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status, number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. Table A.10 shows that the results presented in Table 3 are almost identical if we exclude from our main sample the ten municipalities with more than two hospitals per municipality. The results are also robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are clustered standard errors.

Table A.11: Main results using the weighted score without and with controls

| | **Unhealthy** | **LBW** | **Prematurity** | **Apgar < 7** |
|---|---|---|---|---|
| | Average Health Scores | | | |
| | (1) | (2) | (3) | (4) |
| | **Panel A. Without controls** | | | |
| Coefficient | -0.0064*** | -0.0034* | -0.0037** | -0.0029** |
| Stand. Err. | (0.0022) | (0.0018) | (0.0017) | (0.0013) |
| Adjusted Coeff. | -6.66% | -7.78% | -8.80% | -7.63% |
| | **Panel B. With controls** | | | |
| Coefficient | -0.0057*** | -0.0035** | -0.0032** | -0.0022* |
| Standard Error | (0.0017) | (0.0014) | (0.0013) | (0.0012) |
| Adjusted Coeff. | -5.75% | -7.57% | -8.01% | -5.62% |
| Average Dependent Variable | 0.096 | 0.043 | 0.042 | 0.037 |
| Number of Observations | 237,082 | | | |

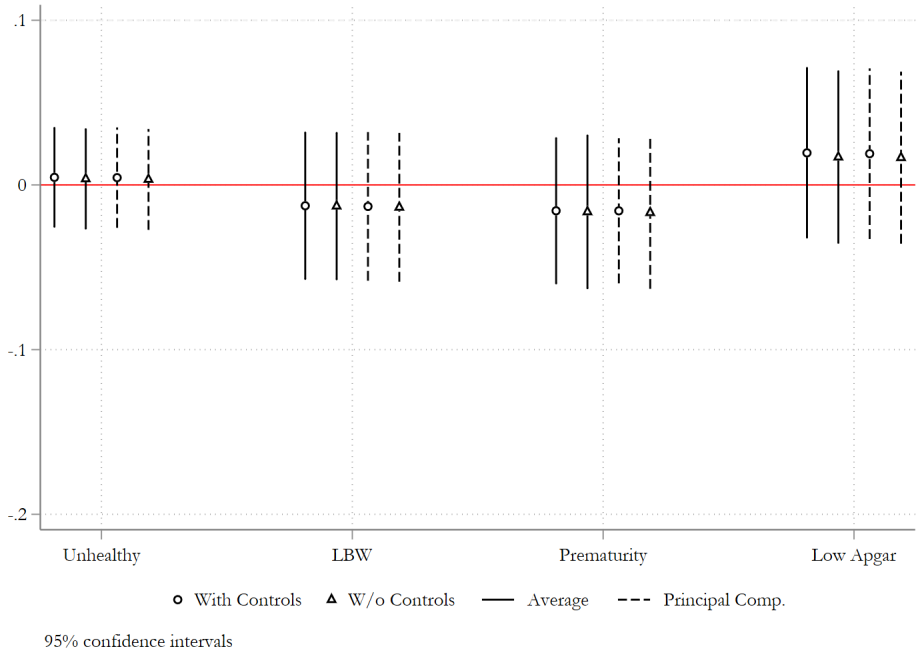
Notes: Table A.11 presents the main results using weighted score. The coefficients represent the effect of an increase of one standard deviation of the physicians' skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status, number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. The table shows that the results are very similar when the weighted score is used as a proxy of physicians' skills. The results are also robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

## Table A.12: Placebo other years

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

Notes: Table A.12 presents the placebo exercise for the main results. The coefficients represent the effect of an increase of one standard deviation of the physicians' skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status, number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

Figure A.6: Placebo using all samples and average scores




Notes: Figure A.6 shows the results of running an exercise analogous to the one presented in Figure 2 but moving the arrival date of the physician four years back (years 2009-2012). The coefficients represent the effect of an increase of one standard deviation of the physicians' 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status, number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to the inclusion/exclusion of controls and the way we measure of skills. These results support the ones presented in Table 5 on the robustness of the estimated zero effect for the placebo tests.

Table A.13: Placebo robustness checks

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

Notes: Table A.13 presents the placebo exercise for the main results. 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. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7, and zero otherwise, low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise, prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise, 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 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, 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, 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, and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. Note that the results are robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are clustered standard errors. * significant 10%, ** significant 5%, *** significant 1%

Table A.14: Placebo estimating a Logit model

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

Notes: Table A.14 presents the placebo exercise for the main results using a Logit. The coefficients represent the average marginal effect of the physicians' skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise; low birth weight is a binary variable that takes the value of 1 if the newborn has low birth weight and zero otherwise; prematurity is a binary variable that takes the value of 1 if the newborn is premature (fewer than 37 weeks of gestation) and zero otherwise; Apgar is a binary variable that takes the value of 1 if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status, number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. Note that the results are robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

Table A.15: Other heterogeneous effects

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

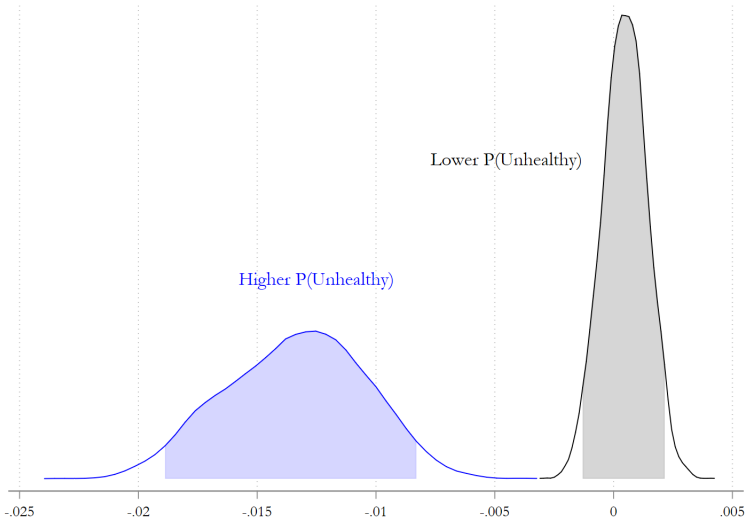
Notes: Table A.15 presents the main estimates by mother and gender of the newborn heterogeneous effects. The coefficients represent the effect of an increase of one standard deviation of the physicians' skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the gender of the newborn; an indicator variable that takes the value of 1 if the mother has at least secondary education and zero otherwise; an indicator variable that takes the value of 1 if the mother is adolescent and zero otherwise; marital status, number of inhabitants in the municipality; number of hospitals per municipality; area; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. These results show that the estimated effects are robust to the inclusion/exclusion of controls and the way we measure of skills. Numbers in parentheses are clustered standard errors.
* Significant 10%, ** significant 5%, *** significant 1%

Table A.16: Antenatal consultations < 4

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

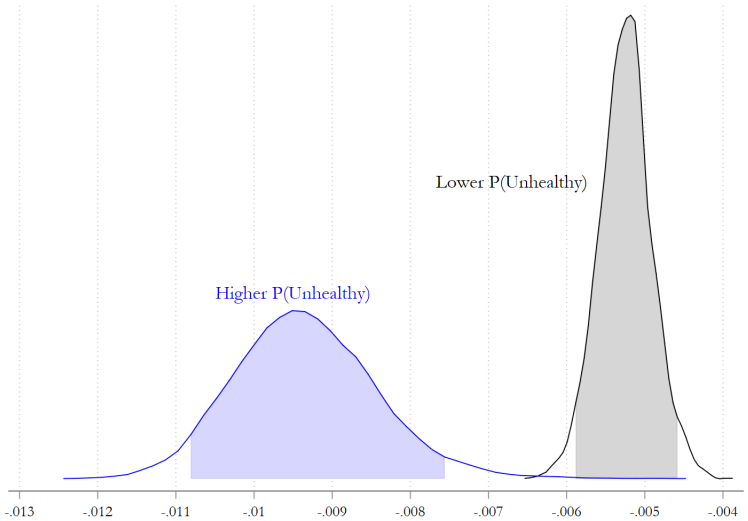
Notes: Table A.16 presents the results for antenatal consultations. The coefficients represent the effect of an increase of one standard deviation of the physicians' skill measure. Relative (percent) effects are computed as the coefficient divided by the average of the dependent variable. Antenatal consultations takes value one if the mother attended to less than 4 consultations while pregnant, an zero otherwise. All regressions control for draw-state fixed effects. Regressions for the coefficients labeled as "With controls" also include the following controls: an indicator variable for the 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; 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; 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; and an indicator variable that takes the value of 1 if the hospital is above the 75th percentile of the distribution of Apgar 1 measured in 2010-2012 and zero otherwise. Note that the results are robust to the inclusion/exclusion of controls and how we measure skills. Numbers in parentheses are clustered standard errors. * Significant 10%, ** significant 5%, *** significant 1%

Figure A.7: Distribution of logit simulations on antenatal consultations by predicted probability of unhealthy newborn




Notes: Figure A.7 presents the distribution of logit simulations on antenatal consultations by predicted probability of an unhealthy newborn. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise.

Figure A.8: Distribution of logit simulations on the probability of being born unhealthy by the (ex-ante) predicted probability of an unhealthy newborn




Notes: Figure A.8 presents the distribution of logit simulations on the main outcomes by predicted probability of an unhealthy newborn. Unhealthy is a binary variable that takes the value of 1 if the newborn has low birth weight or if the newborn is premature (fewer than 37 weeks of gestation) or if the Apgar 1 score of the newborn is lower than 7 and zero otherwise.