README for manuscript AEJApp-2024-0092 "Luck of the Draw: The Causal Effect of Physicians on Birth Outcomes"

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

1 Overview

This document describes the instructions foraccessing the datasets used in the manuscript, and the program files used to replicate the analysis of this study. The program files replicate not just the tables and figures in the main paper, but also all tables and figures in the online appendices. The replication package contains 30 Stata and R scripts, which take about 25 hours to run.

2 Data Availability and Provenance Statements

To replicate the results in the paper, the following databases are required: (i) SSO reports, (ii) Saber PRO, (iii) Vital Statistic Records, (iv) National Census, and (v) ReTHUS. These data files were examined on a secure workstation at the Central Bank of Colombia (Banco de la República de Colombia).

The authors of the manuscript have legitimate access to and permission to use the data used in this manuscript. The datasets used in this study were accessed through the researchers' affiliation with the Central Bank of Colombia. As part of this affiliation, researchers are granted access to data collected by the Central Bank of Colombia and were provided by several institutions mentioned below. Such confidential administrative datasets are not publicly available.

For each dataset, we mentioned what are the institutions and positions within them, that interested researchers should get in touch with to request access to their data. We would be

happy to assist other researchers, who have obtained access to the data, use our programs to replicate all our empirical results. Please note that obtaining access to the databases can take many months. It took the researchers on this project to compile the data over many years.

2.1 Primary Data Sources

- **SSO reports, Ministry of Health:** These are the reports written and published by the Ministry of Health for each state-level draw for the draws conducted in January, April, July, and October 2013, as well as January, April, and July 2014. This dataset includes individual identifiers, the draw date, the state to which each physician applied, whether the physician was selected by the lottery, and crucially, the LHC (Local Health Center) to which each physician was randomly assigned, along with the proposed start date. The research team first accessed this dataset in June 2018. The original data can be accessed via the Ministry of Health. The data can be obtained by accessing the following link:
  https://www.minsalud.gov.co/Paginas/
  Videoconferencia-asignacion-plazas-servicio-social-obligatorio-2013.aspx.
  To cite these data, please use the following citation:
  Ministry of Health (2014). Reports of professionals registered and assigned to the process of assigning places in the mandatory social service.

- **Saber PRO, ICFES:** This dataset, administered by the Colombian Institute for Educational Evaluation (Instituto Colombiano para la Evaluación de la Educación, or ICFES), contains detailed information on physicians' individual performance in the SABER PRO exam, a mandatory graduation test for all professionals in Colombia. It includes test scores from 2010 to 2018 for two health-related modules—Health Care and Disease Prevention—as well as modules on Critical Reading and Quantitative Reasoning. Additionally, the dataset provides rich sociodemographic information about each physician. The research team first accessed this dataset in May 2019. The original data can be accessed via the ICFES. The data can be obtained by accessing the following link:
  https://www.icfes.gov.co/evaluaciones-icfes/resultados/.
  To cite these data, please use the following citation:
  Colombian Institute for Educational Evaluation (2014). Quality evaluation of higher education.

- **Vital Statistic Records, DANE:** This dataset, collected by the Administrative Department of Statistics (Departamento Administrativo Nacional de Estadística, or DANE), contains comprehensive information from all birth certificates filed in Local Health Centers (LHCs) across Colombia's 1,120 municipalities from 1998 to 2018. It provides detailed data on key aspects of each birth, including the birth date, gestational age, health indicators at birth, and the demographic characteristics of the parents. This rich dataset enables the identification of children born between 2013 and 2016 who were exposed to specific teams of physicians, offering valuable insights into the population-level impacts of healthcare interventions. The research team first accessed this dataset in July 2020.
  To access a public version of the Vital Statistic records, visit the DANE website at:
  https://microdatos.dane.gov.co/index.php/catalog/DEM-Microdatos.
  For an extended version of the Vital Statistic records, researchers can utilize the SISPRO Cube provided by the Colombian Ministry of Health. Interested researchers can request access through the official SISPRO portal at:
  https://www.minsalud.gov.co/proteccionsocial/Paginas/SistemaIntegraldeInformaciÃşnSISPRO.aspx.
  To cite these data, please use the following citation:
  Administrative Department of Statistics (2018). Vital statistics records. https://www.dane.gov.co

- **2005 National Census, DANE:** This dataset, collected by the Administrative Department of Statistics (DANE), provides municipality-level population data and additional variables from the 2005 National Census. The information is used to test the randomization of the program and serves as a source of control variables in robustness checks, offering valuable contextual data for the analysis. The research team first accessed this dataset in October 2020. To access a public version of the 2005 National Census data, visit the DANE website:
  https://microdatos.dane.gov.co/index.php/catalog/DEM-Microdatos.
  To cite these data, please use the following citation:
  Administrative Department of Statistics (2005). National Census. https://www.dane.gov.co

- **ReTHUS, Ministry of Health:** The National Registry of Human Resources in Health (Registro Único Nacional del Talento Humano en Salud, or ReTHUS) was created by the Ministry of Health under Law 1164 of 2007, records all individuals authorized to practice

a health profession or occupation in Colombia. This dataset includes detailed information such as the date of degree completion, the date the medical license was granted, and postgraduate qualifications until 2018. The research team first accessed this dataset in July 2022. The original data can be accessed via the Ministry of Health through the following link, or by directly contacting the Director of Human Talent at the Ministry of Health for further assistance:

https://www.sispro.gov.co/central-prestadores-de-servicios/Pages/ReTHUS-Registro-de-Talento-Humano-en-Salud.aspx.

To cite these data, please use the following citation:

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

2.2 Other Data Sources

- **Colombia municipality shapefile:** The research team first accessed this dataset in May 2019. For the original data, visit the DANE website:

  https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/datos-geoestadisticos/.

  To cite these data, please use the following citation:

  Administrative Department of Statistics (2018). National Geostatistical Framework. https://www.dane.gov.co

- **Universities accreditation information:** The research team first accessed this dataset in May 2019. For the original data, visit the Ministry of Education website:

  https://hecaa.mineducacion.gov.co/consultaspublicas/programas. To cite these data, please use the following citation:

  Ministry of Education (2019). National Higher Education Information System.

3 Instructions for Replication

The repository programs are organized into two main folders:

- **Code Data:** This folder contains the scripts required to construct the primary databases

used in the analysis. It includes three Stata scripts that prepare and organize the data for the estimations.

- **Code Results:** This folder includes the scripts necessary to replicate the tables and figures presented in the document and appendix. It comprises 23 Stata scripts and one R script, which handle the generation, storage, and visualization of the estimates.

Additionally, a Stata script named `Do_master` is located outside these two main folders and serves as the central file for executing the analysis. This script sets the directory for the replication folder, specifies the Stata version to be used, and installs the necessary packages to ensure proper program execution. It also automates the sequential execution of all scripts in the repository. When opening the `Do_master` script, the user must update the global root pathname to match the file structure of their system. Additionally, the user needs to modify the directory name in the script to reflect the desired name for their replication folder. These adjustments ensure that the replication process runs correctly within the user's local environment.

4 Programs Description

4.1 Brief guide to program files inside "Code Data"

Scripts are labeled with numbers that represent the steps needed to replicate the main databases without errors. Since the file names are not necessarily self-explanatory, a brief explanation of each script will be provided.

- **Step 1 - Data physician level.do:** Generate and process the variables at the physician level, concerning assignment condition, local health center, starting date, Saber PRO test scores and socioeconomic factors.

- **Step 2 - Final data.do:** Create variables at the newborn and hospital level. Merge with assigned physicians' information, processed in Step 1 and calculate the overlapping periods between newborns and physicians' cohorts.

- **Step 3 - Population at municipality.do:** Obtain the population of the municipalities included in the sample.

4.2 Brief guide to program files inside "Code Results"

The scripts within the folder are organized and labeled according to the corresponding table or figure presented in the manuscript. Each script is named with a sequential number and the title of the table or figure it generates, ensuring clarity and easy reference for replication purposes.

- Figure 1 - Figure A.4 - Heterogeneity in Saber Pro scores.do
- Figure A.1 - Local polynomial regression.do
- Figure A.2 - Heatmap.R
- Figure A.3 - Density physicians.do
- Table 1 - Summary Statistics.do
- Table 2 - Table A.2 - Covariate balance at hospital level.do
- Table 3 - Table A.4 - Figure A.5 - Placebo.do
- Table 4 - Table A.9 - Figure A.8 - Figure 2 - Figure 3 - Main estimates.do
- Table 5 - Table 6 - Heterogeneous effects.do
- Table 7 - Physicians observables and their value added.do
- Table 8 - Table 9 - Figure A.6 - Figure A.7 - Results by predicted probability of unhealthy IV.do
- Table A.1 - Summary Statistics at physician level.do
- Table A.3 - Placebo other leads.do
- Table A.5 - Main estimates Other scores.do
- Table A.6 - Mortality estimates.do
- Table A.7 - Controlling by share of SSO physicians.do
- Table A.8 - Antenatal consultations.do
- Table A.10 - Main results using covariance index.do

- Table A.11 - Main estimates using a Logit model.do
- Table A.12 - Linearity of Main Results.do
- Table A.13 - Main estimates without, with dummy and continuous controls.do
- Table A.14 - Interaction between cohort scores and program scores.do
- Table A.15 - Main results using municipalities with one LHC.do
- Table A.16 - Main estimates using the weighted score.do

5 Process Description Summary

We use reports from the Ministry of Health for the draws conducted in January, April, July, and October 2013, as well as January, April, and July 2014. This dataset includes individual identifiers, the draw date, the state to which each physician applied, whether the physician was selected by the lottery, and crucially, the LHC (Local Health Center) to which each physician was randomly assigned, along with the proposed start date. By using national ID numbers, we link the physicians participating in the SSO program to the ICFES records and retrieve their information from the field-specific medical exams (SABER PRO). We then construct the final dataset, `base_medico_rural_con_fecha_inicio.dta`, which contains each physician's performance in two health-related fields—health care and disease prevention—as well as detailed sociodemographic information and data from each draw.

We use the VSR dataset to create the final dataset, `base_hospital_cohorte_bebe.dta`. By using LHC identification codes, we link physicians to the birth records of the hospitals to which they were assigned. Using the birth date and number of gestation weeks from the VSR, we identify children born between 2013 and 2016 who were exposed to each team of physicians. We also use VSR data from 2009 to 2012 to create mother and hospital-level controls, which help us demonstrate covariate balance at the hospital level and conduct placebo tests.

6 Log Files

The log files for the scripts in both the `Code Data` and `Code Results` folders are generated independently for each script. These log files are automatically created during the execution of the respective scripts and are saved in the designated `Logs` folder. This organization ensures that each log file corresponds directly to the script that produced it, facilitating the traceability and reproducibility of the analysis.

7 Software Requirements

The replication programs were coded using Stata, version 17 and R, version 4.3.1. All the required packages are listed and installed in the `Do_master` script and `Figure A.2 - Heatmap.R` (which is the only program developed in R).

The programs were last run on a 20-core Intel Xeon-based workstation running Windows 10 Pro for Workstations (64-bit), with 9GB of RAM, DirectX 12 support, and 141GB of free space available.

The complete set of scripts in this replication package requires approximately 25 hours to run under the specified computational conditions.