NBER WORKING PAPER SERIES

THE BRAZILIAN BOMBSHELL? THE LONG-TERM IMPACT OF THE 1918 INFLUENZA PANDEMIC THE SOUTH AMERICAN WAY

Amanda Guimbeau
Nidhiya Menon
Aldo Musacchio

Working Paper 26929
http://www.nber.org/papers/w26929

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2020

We thank Marcella Alsan, James Feigenbaum, Carola Frydman, Eric Hilt, Robert Margo, and Edson Severnini, and participants at the NBER-DAE 2019 Summer Institute, the EHA 2019 Meetings, the LACDEV 2019 Conference, the LACEA-LAMES 2019 Conference, Northeastern University, and the Ph.D. seminar series at Brandeis for comments and suggestions. We thank Dani Castillo, Andre Lanza, Pedro Makhoul, Stephanie Orlic, and Uros Randelovic for excellent research assistance. This project was made possible by a Provost Research Grant and funds from the Brazil Initiative at Brandeis University. The usual disclaimer applies. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Brazilian Bombshell? The Long-Term Impact of the 1918 Influenza Pandemic the South American Way

Amanda Guimbeau, Nidhiya Menon, and Aldo Musacchio

NBER Working Paper No. 26929
April 2020
JEL No. I15,J10,N36,O12

ABSTRACT

We analyze the repercussions of the 1918 Influenza Pandemic on demographic measures, human capital formation, and productivity markers in the state of Sao Paulo, Brazil's financial center and the most populous city in South America today. Leveraging temporal and spatial variation in district-level estimates of influenza-related deaths for the period 1917-1920 combined with a unique database on socio-economic, health and productivity outcomes constructed from historical and contemporary documents for all districts in Sao Paulo, we find that the 1918 Influenza pandemic had significant negative impacts on infant mortality and sex ratios at birth in 1920 (the short-run). We find robust evidence of persistent effects on health, educational attainment and productivity more than twenty years later. Our study highlights the importance of documenting the legacy of historical shocks in understanding the development trajectories of countries over time.

Amanda Guimbeau
Brandeis University
415 South Street MC 032
Waltham, MA 02453
amanda2016@brandeis.edu

Nidhiya Menon
Brandeis University
415 South Street Waltham,
MA 02453
nmenon@brandeis.edu

Aldo Musacchio
Brandeis International Business School
415 South Street MC 032
Waltham, MA 02453
and NBER
aldom@brandeis.edu

1 Introduction

The long reach of history in shaping economic development is well understood. From colonial institutions to slavery, that the past matters in charting the present trajectories of countries is now widely accepted. However, relatively little has been written on the scarring effects of historical health shocks. Although there has been significant interest in the impact of health events since Almond (2006) studied the long-term effects of the greatest epidemic of modern history - the 1918 influenza pandemic - which recently crossed its centennial, we still are not clear how impacts may differ when we look beyond cohort controls and importantly, when we analyze effects in the context of developing countries. This is the gap we address in this paper. The 1918 influenza pandemic, by its sheer magnitude and features, provides a unique natural experiment to test a range of hypotheses related to the short and long-run consequences of exposure to diseases, and offers an interesting framework to study the effects of an extraordinary mortality shock on demographic and other outcomes.

We study repercussions of the 1918 "Spanish Flu" on demographic, human capital and productivity outcomes in the State of Sao Paulo, Brazil, in the short- and long-runs. Although today Sao Paulo is Brazil's financial center and South America's most populous city, in the early twentieth century, it was far from such. Given the lack of resources for remedial action and the relatively more primitive health care infrastructure, it is likely that the pandemic's immediate and lingering effects were more harmful in poorer nations like Brazil that were marked by social inequalities and a nascent health system. Moreover, the pandemic coincided with the ending years of the First World War. It is thus hard to disentangle the detrimental impacts of the 1918 flu from the widespread destruction caused by the War, particularly when we focus on the main actors involved such as the United States. Brazil's contribution to the Allied war effort began in 1917 and was minimal, and the country did not experience the level of destruction that those in the North endured. The First World War is thus less of a confounding factor in the historical Brazilian context.

Records reveal that in Sao Paulo (city), the disease caused 5,331 deaths in the short period between mid-October and mid-December 1918 (Massad et al., 2007), and infected up to 350,000 people, two-thirds of the population of Sao Paulo (city) (Bassanezi, 2013, Barata, 2000, Bertolli, 2003). As we show below, the pandemic's duration and intensity differed significantly across geographical markers of districts (the spatial unit of our analysis) in the State of Sao Paulo. We exploit this spatial variation to link the number of influenza-related deaths to a range of outcome variables over time. We accomplish this using detailed district-level historical data on vital statistics, health, education, and productivity variables. We complement this data with information drawn from official statistical reports on the pre-existing human capital framework that was in place at the time the disease arrived in Sao Paulo.

Our study considers two time horizons: 1920 (the short-run) and 1940 (the long-run). The primary sources used are the Brazilian censuses from 1920 and 1940. Using two stage least squares methods, we instrument for respiratory deaths normalized per 1000 people, our proxy for influenza-related deaths during the pandemic, with average temperature and rainfall in October. The "Spanish Flu" arrived in Sao Paulo in late September 1918 and as we discuss in detail below, air temperature (sunlight) reduces the incidence of influenza whereas rainfall, by increasing humidity, has the opposite effect. We find that as of 1920, infant mortality and still births increased whereas sex ratios at birth declined. This is entirely consistent with expectations as fetuses and infants are particularly susceptible to health shocks. Moreover, male fetuses are relatively more vulnerable than female fetuses, leading to excess male mortality in response to negative health shocks of this nature (Sanders and Stoecker, 2015). Furthermore, we find that short-run agricultural productivity, as measured by the volume of coffee, rice and maize per capita, declined in 1920.

Turning next to the long-run, we consider impacts on literacy rates in 1940, one of the canonical indicators of human capital. We begin by estimating effects on two different age groups (20-29 and 30-39 years), positing that the cohort aged 20-29 years was directly exposed to the pandemic (either *in utero* or at a very young age) while the other cohort was not. In sum, our findings indicate that the pandemic led to the deterioration of literacy, especially for women. This is consistent with other results that show that female educational levels are relatively more affected after natural disasters (Neumayer and Plumper, 2007, Caruso, 2015, Caruso and Miller, 2015). Alternatively, we find an increase in male literacy rates for the cohort that was directly exposed to the pandemic, consistent with positive selection resulting from excess male mortality (which is in accordance with existing empirical findings that the long-term effects of early-life shocks involve boys' culling and girls' scarring). Further, these results are broadly consistent with gender-disaggregated results from the 1940 census that do not demarcate cohorts, and with methods that use cohort-level variation using the 1960, 1970 and 1980 Brazilian censuses in a framework similar to Almond (2006) and Beach et al. (2018). In terms of health, respiratory deaths from the pandemic have a significant impact on the normalized number of inpatient hospital admissions in 1940. We find that productivity was also affected - the primary sector's output per employee and per establishment declined as of 1940.

Our paper contributes to research that documents the path-dependency of human capital, inequality, poverty and development in response to historical epidemics and shocks (Alsan, 2015, Bleakley, 2010, Clay et al., 2018). This literature focuses on the deep roots of economic development and seeks to understand how past episodes explain variations in contemporaneous growth rates, stocks of civil/social capital and political outcomes. However unlike previous work, our study benefits from the fact that Brazil was not a major actor in the First World War, and from the fact that given its location in the Southern Hemisphere, the 1918 Pandemic arrived in Sao

Paulo in the Spring. These factors make the context of our study unique, and allow us to distill results from an environment that is cleaner in many ways from other research that has concentrated on the developed world. To the best of our knowledge, our paper is the first study of the 1918 pandemic's impact on a range of demographic, human capital and productivity outcomes in a developing country that uses both cohort and geographical variation with rich contemporary controls for geography, immigration, health and sanitation to investigate the ways in which this unanticipated and intense health shock changed the trajectory of development for a major location in Brazil. The highly disaggregated nature of our historical climate data further allows us to design a strategy that overcomes endogeneity concerns. We extract the relevant data using latitude and longitude coordinates from the Climate Change Knowledge Portal (World Bank Group), and assemble a novel dataset that combines geographical conditions with archival micro-data on the spatial and temporal distribution of respiratory deaths and historical institutions and context during the pandemic years. Our use of information at such a granular level to compile a rich database is a strength of our paper that is a significant improvement on existing research. Our results indicate impacts that are on average of the same sign but larger than those estimated on comparable outcomes in developed countries. An important exception is our gender-disaggregated results on education which indicate an increase in male literacy in the immediate aftermath of the pandemic and twenty years later. We discuss how positive selection and segmented labor markets may be explanations for this result which is not in keeping with other studies that have considered the immediate and long-term human capital consequences of the 1918 pandemic. Taken together, our research highlights more comprehensively than other studies what economic historians mean when they say that the 1918 influenza pandemic forever changed the social, economic and cultural landscape of countries.

The paper is structured as follows: Section 2 discusses relevant literature and Section 3 describes the historical background of the pandemic in Sao Paulo, Brazil. Section 4 describes the data and provides the summary statistics on vital, geographical, demographic and economic characteristics. Section 5 describes the empirical methodology, and Section 6 details results from the short-term analysis. Sections 7 outlines results from two decades after the event, and Section 8 describes robustness and falsification tests. Section 9 concludes.

2 Literature

Several studies explore the historical roots of comparative development and use major demographic events, geopolitical turning points, or changes in the institutional environment, to explain existing disparities between nations and in-country variation in economic outcomes. [1] Acemoglu

[1] These include Huillery (2009) which investigates the long-term impact of colonial public investments in French West Africa; Nunn and Wantchekon (2011) that traces mistrust to the slave trade in Africa; Putterman and Weil

et al. (2001) demonstrates that the disease environment faced by European settlers matters for institutional legacies and thus, for long-term economic development. Beach and Hanlon (2017) using wind patterns for identification, find that British industrial coal use in 1851-60 had significant mortality effects. These results provide further support for findings in Hanlon (2019) that long-run city growth in Britain in the nineteenth and twentieth centuries was negatively affected by local industrial coal use.

Looking specifically at the pandemic of 1918, Almond (2006) evaluates empirically whether exposure to this health event had repercussions later in life. Using US census microdata from 1960-1980, the study finds that cohorts that were *in utero* during the pandemic have lower socioeconomic status and lower levels of education, health, and reduced employment outcomes when seen in the future. These results support the *fetal origins hypothesis*[2]. In a related paper, Almond and Mazumder (2005), using data from the Survey of Income and Program Participation (SIPP) from 1984-1996, find that birth cohorts *in utero* during the 1918 Pandemic have poorer health outcomes almost 65-80 years after the event. Beach et al. (2018) use linked data from WWII enlistment records and federal censuses, and an individual-level panel dataset in a framework that accounts for fixed-effects, observed parental characteristics, and a wide set of controls to find that that the evolution of human capital was impaired by the pandemic. Using data from Taiwan, Lin and Liu (2014) conclude that the 1918 birth cohorts are shorter as teenagers, have less education, and are more prone to various diseases. These results are in line with those in Chul Hong and Yun (2017) which uses data from colonial Korea. Richter and Robling (2013) and Cook et al. (2018) are other studies that evaluate the impact of the Spanish Flu on multiple generations.

Studies that measure the aggregate effects of the pandemic on earnings, human capital accumulation and economic growth include Karlsson et al. (2014) which finds that regions more burdened by the pandemic in Sweden had higher future poverty rates, and Percoco (2016) which uses mortality rates across Italian regions to find that exposure to the flu lowered educational attainments for those who were *in utero* and in early childhood. Donaldson and Keniston (2016) analyze the impact of the event in colonial India and obtain results consistent with Malthusian growth theories - Indian districts heavily burdened had higher fertility rates in the aftermath of the shock, and there is also evidence of increased investments in child quality (literacy rates for men/boys rose). Bakken and Husoy (2016) match 1912-1920 data on influenza mortality and other demographic estimates to the Norwegian 1960 census and find that exposure to prenatal influenza lead to significant declines in the years of education for men, with a larger effect in the poorest municipalities. Considering Brazil specifically, Nelson (2010) examines the impact of the pandemic on later life

(2010) which studies the role of human capital in long-run development; and Dell (2010) and Glaeser et al. (2004). Guiso et al. (2016) shows that Italian cities with self-government in the Middle Ages have relatively high levels of civic capital today, and Rocha et al. (2017) exploits variation induced by the state-sponsored settlement policy during the historical episode of mass migration in Brazil to evaluate path dependency in human capital formation.

[2]The hypothesis that insults to the developing fetus, nutritional and otherwise, have lingering effects; attributed to Barker (1995).

outcomes (education, employment and wages) using labor surveys from 1986 to 1998. In keeping with previous work, those who were immediately impacted by this mortality shock suffered in the long-run along these dimensions.

Our study complements the research discussed above by evaluating a whole spectrum of outcomes that span demographic measures, educational outcomes, health variables and productivity measures, over different time-lines. Importantly, we are able to control for contemporaneous socio-economic, infrastructural and regional characteristics from the time the pandemic arrived in Sao Paulo, as well as a comprehensive set of controls from 1872 that help us account for possible pre-trends in the data. Previous data limitations have meant that studies have been able to consider only a few outcomes and controls at best. We are more fortunate in this regard. The disaggregated nature of our historical data matched with censuses from 1920 and 1940 enables us to analyze the consequences of the 1918 influenza pandemic in Brazil in an exhaustive manner.

3 Historical Background

Sometimes described as the "greatest medical pandemic of modern times", the 1918 influenza pandemic resulted in global mortality estimates ranging from 20 to 100 million within the span of a few months (Alonso et al., 2016). The pandemic reached Brazil in the beginning of Spring, on 14th September 1918. *Demerara*-an English-flagged ship, entered the port of Recife, a northern Brazilian city, and then anchored in the harbor city of Santos, Sao Paulo (Massad et al., 2007). The sailors on board *Demerara* were sick since they had been directly exposed in Dakar, Senegal, a stop on their way back from Europe (Alonso et al., 2016). A few days after the arrival of the ship in Santos, the influenza virus arrived in the interior parts of the country and many cases were reported in Rio de Janeiro and Sao Paulo, and in other cities in the northeast. The epidemic's pace and intensity could not be contained and it spread rapidly to the interior districts.

Evidence suggests that Brazilian authorities did not anticipate the deadly effects of the pandemic. The information that reached the country warning about the severity of the flu overseas did not receive due importance. Hence when it arrived, it easily overwhelmed public authorities. The few measures in place proved useless when confronted with the outbreak of a severe public health crisis, as seen by newspapers extracts from the time (Figure 1 and Figure 2). Prices of foodstuff (including milk, meat, and lemons) and medicines (mostly quinine-considered to be a powerful drug for any illness) quickly surged as shortages set in (Hochman, 2016). Improvised hospitals and healthcare posts had to be established in many cities that were trying to keep up with rising death rates. Sao Paulo (city) reported 116,771 cases of influenza, a prevalence rate of 22.32% (Nelson, 2010), for an estimated population of 523,194 inhabitants, and 5,331 deaths in the short period between mid-October and mid-December 1918 (Massad et al., 2007). The Brazilian president

Francisco de Paula Rodrigues Alves, newly elected for a second term in 1918, succumbed to the flu in January 1919. Some studies estimate that approximately 350,000 (two-thirds of Sao Paulo's population) might have been infected (Bassanezi, 2013, Barata, 2000, Bertolli, 2003). Our data suggest that in 1918 and 1919, when the pandemic was at its peak, the percentage of deaths attributed to influenza-related causes stood at 52.6% and 53.8%, respectively.

To keep these numbers in perspective, Pennsylvania, Maryland and Colorado in the US, with influenza deaths of 883.1, 803.6 and 776.5 per 100,000 of the population, respectively, had some of the highest mortality rates in 1918.[3] Garrett (2007) explains that the first wave in the US occurred in March 1918 and lasted through summer of 1918. The second wave in Fall of 1918 was worse. Unlike the US, there was only one wave of the pandemic in Brazil. In Figure 3, we show the scale of respiratory deaths per thousand people for the state of Sao Paulo relative to that of US cities in 1918. [4]

The municipality Capital (in Sao Paulo), for which monthly data is available from 1917-1920 from Annual Statistical Reports, provides a snapshot of the situation[5]. Figure 4 shows mortality patterns in this region. Compared to the same months in 1917, there were 514 and 5,274 more deaths in October and November 1918, respectively; with the total number of deaths up by 87.3% in this municipality. In Figure 5, we show the monthly total deaths per 100,000 of the population from January 1917 to December 1919. The mortality shock due to the pandemic is evident. Figure 6 demonstrates that both neo-natal and post-natal mortality (measured as deaths per 1000 live births) also peaked in the Capital in October 1918. In Figure 7, we aggregate the district-level data up to the municipality level for the period 1917-1920, and show the spatial variation in influenza-related death rates using the 1920 boundaries for the state of Sao Paulo.

The pandemic resulted in a higher mortality rate among prime-age adults, creating a 'W'-shaped age-profile distribution of affected groups [6]. A reason proposed for this pattern is that young male adults in particular might have been more vulnerable given their relatively higher participation rates in the labor force, and their disproportionate exposure to other diseases such as tuberculosis. Another explanation from the medical literature is that the "cytokine storm," that causes an

[3]Mortality rates reported in (Garrett, 2007) are drawn from *Mortality Statistics in 1920* and include influenza and mortality estimates.

[4] Crosby (1989) indicates that the highest mortality rate in the US was more than 10 out of 1000 people in New York city. Our data indicates that in Sao Paulo, the comparable estimate is almost 10 out of 1000 people (when deaths due to unknown causes are also considered). The average for US cities stood at 5.8 out of 1000. (Clay et al., 2019) notes that variation in mortality across U.S. cities were related to several factors including pre-pandemic levels of infant mortality, illiteracy and air-pollution.

[5]The Capital area is representative of the State of Sao Paulo for the purpose of our analysis and includes the following sub-regions: Se, Mooca, Consolacao, Bom Retiro, Cambuci, Santa Cecilia, Perdizes, Bela Vista, Vila Mariana, Bras, Penha de Franca, Ipiranga, Santana, Lapa, Nossa Senhora do O, Sao Miguel, Butanta, Osasco, Liberdade, Belenzinho, Santa Efigenia.

[6]The 1918 Flu killed relatively younger men and women aged 15-44, unlike other influenza epidemics that typically kill mostly young children and older cohorts resulting in a U-shaped distribution.

excess production of immune cells and their related compounds, cytokines, is evident in healthy young adults aged 20-40 years old during a flu infection, and this strong immune reaction results in premature death (Loo and Gale, 2007, Kobasa et al., 2004). Children and seniors with weaker immune systems are less affected as the risk is lower that their immune systems will overreact.[7] Given its sudden, unanticipated arrival in Spring in the Southern Hemisphere, and with its short duration of about ten weeks, existing studies posit random assignment of the infection in Brazil (Nelson, 2010).

In response to the onset of the pandemic, public health authorities designed campaigns to disseminate information on preventive and curative measures.[8] These included isolation, good personal hygiene, no extra work that could result in extreme fatigue, and quarantine measures in several areas. Furthermore, there was overall confusion as statements released by medical establishments that conveyed that the epidemic was mostly benign was in dissonance with published data that indicated an exponential growth in the number of victims. Hospitals quickly ran short of medicines and a full blown public health crisis ensued in Sao Paulo as Brazilian authorities struggled to keep pace.

4 Data and Summary Statistics

Using historical and archival records, we construct a unique database on health outcomes related to major disease categories and socioeconomic indicators for districts in the state of Sao Paulo in the early decades of the twentieth century. We complement the data on district-level deaths by cause with measures of infrastructure, demographic information and geographic variables. We accomplish this by digitizing statistical reports and by matching these to Brazilian regional census data from 1920 onwards to obtain a panel of information. Our spatial unit of analysis is at the district level-a disaggregated territorial stratification of the state of Sao Paulo.[9] Use of districts gives us 350 unique geographical observations at the district-level for the period 1912-1921, which we combine with other historical and contemporary data from 1872, 1912-1921, and 1940.

An important consideration while creating our dataset is ensuring that standardized boundaries are tracked over time. The boundaries of many districts and municipalities changed considerably

[7] Unfortunately, due to data limitations, we are unable to compare prevalence rates across different age groups.

[8] Failure to comply with these measures may explain why different areas in Sao Paulo were impacted differently by the virus. For instance, Bassanezi (2013) reveals that two municipalities, Campinas and Riberao Preto, that had suffered from yellow fever in the late nineteenth and early twentieth century, acted on these measures propitiously (Bertucci-Martins, 2005). Sorocaba, relatively smaller than Campinas and Riberao Preto, and with less experience in epidemics, struggled more as they failed to adopt the right initiatives.

[9] A district in Brazil is an administrative unit within a municipality. We choose to use district-level data rather than municipality-level data for three reasons. First, the data on vital statistics for 1917-1921 is available at the district-level, and second, using districts provides a larger sample size. Third, use of districts allows a more refined analysis of realities at a very microeconomic level.

over the time period of our analysis.[10] We use official territorial and administrative maps that provide information on the evolution of districts over time to match the data in a consistent manner.

We begin by describing the key variables of our analysis, and the source(s) from which they are obtained. Table A4 in the Appendix provides details on other variables used in this study.

4.1 Deaths by Cause and Vital Statistics

The Sanitary Authority, created in 1892, was responsible for ensuring the collection of reliable health statistics in Sao Paulo. These statistics are publicly available from 1901-1928 (Bassanezi, 2013). The Annual Statistical Reports of Sao Paulo contain the vital statistics required for our analysis. More specifically, we obtain deaths by cause for 14 major disease categories, allowing us to construct our key variable of interest: deaths rates from respiratory infections from 1912-1921, our best proxy for influenza-related deaths.[11] The Annual Statistical Report also provides information on economic and financial statistics including municipalities' receipts and expenditures, which we use as controls in some of our models. We also obtain the initial shares of influenza-related deaths for 1915 in the *Annual Statistical Report of Sao Paulo (1915)*, for which deaths by cause are available for 180 municipalities only. Further, we use the *Demographic Studies: The Population of Sao Paulo in the last decade: 1907-1916* to obtain pre-pandemic population statistics for most municipalities. The nationwide influenza-related deaths are constructed from the yearly statistical reports of 1917-1921.

We use the in-patient hospital and asylum admissions rates in 1940 (obtained from *Annual Statistical Report of 1940*) as a medium run measure of health. We use per capita municipality expenditures on hospitals and health from the same source as additional controls.

4.2 Climate Data

The Annual Statistical Reports of Sao Paulo provide few meteorological observations. For instance, the 1917 report provides monthly temperature for less than 65 regions and the monthly precipitation for less than 35 municipalities. Atmospheric pressure, humidity level and detailed air temperature-related statistics are only available at the aggregate level. Given these limitations, we compile our climate data using averages from the World Bank Climate Change Knowledge Database, where we extract relevant data on temperature and rainfall in October from 1901-1930,

[10] The first census of 1872 contains data on 88 municipalities; by the 1940 census reports, the number of municipalities reached 350.

[11] For reasons explained in the next section, we also collect data on other vital statistics. These data are further complemented with information from a special official Sanitary Report in 1918.

using latitude and longitude coordinates of districts (these data are not available for October on an annual basis)[12]. However, models include district-level averages for temperature and rainfall. This means that we use deviations in these measures from their trend; but these deviations are of a more aggregate nature given the data constraints we face. Further, we collect long-term averages for the period 1901-2016 and use deviations of October's average (over 1901-1930) temperature and precipitation from these long-term trends as a robustness check. These results are reported in Table A6 and show that although these deviations have the expected sign, they have low predictive power (the F-statistic is smaller than 10). We also carry out other falsification tests as detailed in the next section.

4.3 Sanitary and Health Infrastructure

We collect information on the number of doctors, chemists and midwives, and the number of people with mental and physical deficiencies (per 1000 inhabitants) as of 1872, and match these to the 1920 district-level data for use as proxies for historical health indicators in a period before the pandemic struck. Information on water and sewage systems is obtained from the *Sanitary Service Report of the State of Sao Paulo: Annual Demographic and Sanitary Statistics Section (1920).* This report is also used to obtain the number of hospitals and old-age homes and specialized maternity hospitals for use as controls when the outcomes are child death, or still births, or deaths caused by specific diseases. The *1910 Annual Statistical Report of Sao Paulo (Volume II)* is used to obtain the pre-pandemic public expenditures on cleaning, waste disposal, and maintenance of sanitary conditions in 132 municipalities.

4.4 Geography and Railroads

We control for altitude, latitude and longitude in our models. Altitude and related data are collected from the different volumes of *The Encyclopedia of Brazil Municipalities (IBGE 1957,1958)* for Sao Paulo.

The dummy for the presence of railway in the district is created from the *Secretariat Report of Agriculture, Commerce and Public Works of the State of Sao Paulo; Coffee: Statistics of Production and Commerce (1920).* We know the number of railway companies and stations in 1920, and the year in which the railway network was established. Distance to capital (both in a

[12]This is an online platform provided by the World Bank Group that provides access to climate change data. Amongst many other services, it provides spatial and temporal averages of rainfall and temperature for the periods 1901-1930, 1931-1960, 1961-1990, 1991-2016, and 1901-2016. As acknowledged on the website, the temperature and precipitation data is obtained from the Climate Research Unit (CRU) of the University of East Anglia, reprocessed by the National Center for Atmospheric Research (NCAR) of the University Corporation for Atmospheric Research (UCAR).

straight line and more precisely) comes from the same 1920 report and from the *Ipeadata* (the Institute of Applied Economic Research in Brazil).

4.5 Productivity

The *1920 Census of Agricultural and Industrial activities* is the source for our productivity measures in 1920. We calculate productivity measures at the municipality level by combining data on establishment size (measured by the number of employees and by size), production of different commodities, and the characteristics of small scale and large scale business from the 1940 census for Sao Paulo.

Data on the number of people employed in schooling activities (school workforce) is obtained from the *1940 census (Parte XVII – Sao Paulo, Volume 1)*. The *Annual Statistical Report of 1940* is used to obtain the total municipality expenditure on education[13], and the budget share that accrues to school supplies, materials and to teachers.

The 1940 census (*Parte XVII – Sao Paulo, Volume 3*) is used to obtain total expenditure per establishment for the primary sector (and for the subsectors agriculture, farming and livestock), and the number of plows used per establishment, which is a proxy for capital utilization/mechanization. We also compute expenditures per agricultural establishment on new seeds, fertilizers and insecticides, and on salaries and the acquisition of new machinery and animals in 1939. These variables are used as controls in the agricultural productivity models.

4.6 Census Data

We use the 1872, 1920, and 1940 censuses for the State of Sao Paulo to estimate the short-term and long-term effects. The first round of 1872 allows us to obtain pre-pandemic initial conditions. In addition to those noted above, other variables that we obtain from this Census include the share of foreigners, the share of literate people, the share of people who were slaves, population density, and race. Controling for initial conditions is important to ensure that we have a baseline for the degree of development, urbanization, infrastructure, social aspects, health-service, and the sanitary environment before the Flu Pandemic struck (that is, these variables help to control for pre-trends that may exist in the data).[14]. The 1920 census is used to compute demographic and economic variables.[15] The 1940 census is used to construct longer-run measures, as has been

[13]These data allow us to calculate the share of total municipality expenditures devoted to the Secretary of Education and Public Health.

[14]Note that we use these measures in proportions.

[15]Amongst others, we calculate sex ratios for Brazilians and foreigners, and literacy rates by age groups and gender.

detailed above.

4.7 Summary Statistics

Tables 1 and 2 present summary statistics for the dependent variables and selected explanatory variables in our analysis. Variables in Panel A of Table 1 are the dependent variables in the short-run of the pandemic and include infant mortality, still births, the literacy rate for males and females, the sex ratio at birth, amongst others. Focusing on a few, the mean value for the infant mortality rate is 0.02 and that for still births per total births is 0.05. As expected, male literacy is higher than female literacy. The mean sex ratio at birth in 1920 was 121 (out of 100), with a standard deviation of 82 (out of 100), suggesting that in some districts at least, the levels were far from the natural rate of 105 out of 100. Panel B of Table 1 report statistics for the long-run variables of interest. These include educational measures disaggregated by gender, and health measures including hospital admissions.

Table 2 reports summary statistics for the variables used as controls in the models. These are arranged by panels that focus on deaths by cause, geography, variables from the 1872 census, and other measures. Our key variable of interest is in Panel A of Table 2 - respiratory deaths per 1000 inhabitants - which has a mean of 1.74 and a standard deviation of 2.03 for the period 1917-1920. Average October's temperature was 68.43 Fahrenheit with a standard deviation of 3.76; while average October's precipitation was 124.15 mm (standard deviation 8.39mm). We note the high standard deviation for health and sanitation municipality expenditure in 1910, and the adult sex ratio in 1920 had a mean of 112 (out of 100) with a standard deviation of 59 (out of 100). Overall, the data provides evidence of the relatively low levels of development in most districts during the time period under analysis. Further, these statistics reveal that there was significant geographical variation in these measures, especially in those pertaining to literacy, health and infrastructure.

5 Estimation Methodology

Our methodology follows two broad steps. First we use the 1960-1980 Brazilian Census data from IPUMS to replicate the methodology in Almond (2006). Of course, since these Census data are not as detailed as the US Census, we are limited in the outcomes we can estimate. Moreover, given data limitations, we assume that the current place of residence is the same as their birth place/place of residence when the pandemic arrived in Brazil. There is also some evidence that year of birth may be mis-reported in the 1960 Census. In particular, 1920 is reported as the year of birth by people who may have been born earlier. Since we condition on 1919 similar to Almond (2006), we have a conservative bias in our results (we underestimate effects). Within

this broad framework of using cohorts, we also replicate the methodology of Beach et al. (2018) which combines cohort controls with parental characteristics. Again we are limited in what we can do given the information in the Brazilian Census data. The results from these cohort models are reported in Section A1 of the Appendix and resonate broadly with those from our preferred spcecification.[16]

Second, as we believe that relying on cohorts alone are not sufficient given the dramatic transformation that Brazil was undergoing in these years, we use our rich disaggregated data compiled from various historical and contemporary sources to study the long-term consequences of the pandemic. This is our preferred specification which is described in detail below. Moving beyond the Censuses alone also allow us to analyze a wide range of demographic, human capital and productivity outcomes.

5.1 Empirical Specification

We examine the impact of the 1918 Flu Pandemic using the following model:

$$y_d = \beta_0 + \beta_1 Flu_d + \beta_2 X'_d + \beta_3 X'_{d0} + \beta_4 R + \epsilon_d \tag{1}$$

where y_d is the outcome of interest for district d in either 1920 or 1940. Flu_d is the respiratory death rates in district d (from 1917-1920).[17] X_d is a vector of district-level controls including geographic variables such as altitude, latitude, longitude, and other controls including the distance to capital, a dummy for the presence of a railway, and the interactions of these variables.[18] X_{d0} is a vector of initial conditions obtained from the 1872 census that includes the share of foreigners, the share of literate people, the number of doctors, chemists and midwives, population density, the share of people who were slaves (during the pre-abolishment era), the share of people of different races, and employment in different economic sectors. R denotes region fixed-effects. ϵ_d is the idiosyncratic error term. The coefficient of interest is β_1, the impact of respiratory death rates on the outcome of interest.

There are several reasons why respiratory deaths may not be exogenous. First, measurement error. Official published data may understate true death rates, particularly during the peak of the health crisis between October and December 1918. This may have happened unintentionally as in the confusion that followed the initial outbreak, doctors had little time to make accurate entries. Further, as Richter and Robling (2013) note, panicked and overwhelmed doctors probably chose to make optimal use of their time by treating long lines of frail patients, and focusing on curative

[16] In addition to literacy rates, we estimate models where the outcome is high school completion rates. The latter are not reported in the paper but are available on request.

[17] When we estimate the short-run impact on 1920 variables, the sample is restricted to 1917-1919.

[18] Other controls are added in 1940, depending on the outcome of interest.

work rather than working on long descriptive death reports required by government officials. Others probably struggled to provide the right diagnosis and to assign the cause of death. Figure 8 shows the spatial variation in death rates caused by respiratory infections (left map) and by unknown causes (right map) only in 1918. The maximum value for respiratory deaths per 1000 of the population was 4.5 while that of unknown causes stood at 45. This suggests that many cases were recorded as deaths caused by unknown diseases given the difficulty of recognizing the symptoms and the public health crisis triggered by the pandemic[19].

Second, omitted variables may be simultaneously correlated with respiratory deaths and the outcomes we consider. We control for all that the data allow us to do. In particular, to control for the mortality gradient by socio-economic class, sanitary conditions, literacy, and nutrition, we include controls related to health, demography, human capital, levels of economic development, sanitary infrastructure (including water and sewerage), and the municipality's public health expenditure. Models further include a set of demographic and economic initial conditions from the 1872 census (as in Rocha et al. (2017)) to control for differences in baseline conditions, and for existing pre-pandemic trends.[20]

The natural selection that results from excess male mortality in response to aggregate health shocks of this nature is evident in our study. As we note above, female fetuses are more resilient to shocks *in utero* (Noymer and Garenne, 2000, Hamoudi and Nobles, 2014, Sanders and Stoecker, 2015). We rely on positive selection of male fetuses to explain some of our results that follow, and assume that net of all the controls we include, the distribution of births by gender is not further distorted by unobservables.

5.2 Instrumental Variable

Our preferred specification uses an instrumental variables approach that relies on the fact that seasonal patterns and average environmental conditions observed across regions in the month of October can explain part of the variation in the prevalence of viral respiratory infections and influenza-related diseases.[21] In doing so, we draw on the medical and epidemiological literature

[19] In results not reported, there is evidence that in comparison to 1915, the variance of the distribution of deaths caused by unknown reasons (per 1000) rose between 1917 and 1920, whereas the first two moments of the distribution for respiratory death rates remain mostly the same between 1915 and 1917-1920. The mean for the distribution of unknown death rates also remains about the same between 1915 and 1917-1920. Given that these distribution parameters remain about the same, we conclude that the measurement error is classical in nature.

[20] In Table A3 in the Appendix, we show the sample mean differences for selected geographic, climatic, demographic, economic and health initial characteristics for two sets of regions. These are those with above and below median flu exposure as measured by respiratory death rates between 1917 and 1920. All the variables where these means are statistically different are included in our analyses.

[21] Donaldson and Keniston (2016) use climatic variations in their instrumental variables method and rely on measured absolute humidity while controlling for normal October absolute humidity of the district. As we note above, we are limited in our ability to use climatic variables from October 1918 alone given data availability.

that the incidence of an influenza epidemic depends largely on climatic conditions (Polozov et al., 2008, Tamerius et al., 2013, Slutsky and Zeckhauser, 2018). In particular Slutsky and Zeckhauser (2018) notes that sunlight protects against the flu, implying that there should be negative association between temperature and respiratory deaths. If temperature does not decline beyond a certain level (about 21 Celsius or so), rainfall increases the incidence of the flu (Tamerius et al., 2013). We rely on (only) October's temperature and precipitation that is calculated as the average over 1901-1930. Our regressions also include controls for average temperature and rainfall in the district to ensure that these instruments are plausibly exogenous. We present tests of instrument validity and instrument sensitivity (where the 1901-1930 measure is relative to the 1901-2016 average) below.

Figure 9 shows the spatial variation in average October's temperature and precipitation, respectively. The mean temperature for all districts is 68.43 Fahrenheit and the mean precipitation is 124.15 mm. In order to ensure that these 1901-1930 averages are accurate, we collect data for the few observations available in the Annual Statistical Report of 1917 to obtain a mean of 68.54 Fahrenheit and 119.4 mm for October's temperature and precipitation, respectively, which are close. Panels A and B of Figure 10 portray binscatter plots (with geographical controls) for respiratory death rates and October's rain and temperature. Panel A focuses on the period 1917-1920 while Panel B uses only 1918 data. Both panels show that October's rain and temperature impact predicted respiratory death rates; the empirical results that follow confirm that respiratory deaths are inversely related to temperature and positively related to precipitation.

The first stage regression for respiratory death rates is:

$$Flu_d = \alpha_o + \alpha_1 OctTemp_d + \alpha_2 OctRain_d + \alpha_3 Geog_d^{'} + \alpha_4 Ini72_d + \alpha_5 R + u_d \tag{2}$$

where $OctTemp_d$ and $OctRain_d$ denote average October's temperature and precipitation for each district; $Geog_d$ is a vector of geographical controls including average temperature and rainfall in the district, and $Ini72_d$ is a vector of baseline socioeconomic characteristics. R are region dummies, and u_d is the standard idiosyncratic error term. Results for the first stage regressions are shown in Table 3.

Table 3 (columns (1)-(3), for our period 1917-1920) reports that temperature has a negative effect while precipitation has a positive effect on respiratory death rates. In columns (2) and (3), the results remain unaltered with the inclusion of district-level controls. The estimates on temperature and rainfall (along with the full set of controls) explain about 44.3% of the variation in respiratory death rates. Moreover, the F-statistic on identifying instruments are above the rule-of-thumb threshold value of 10. To ensure that our instruments affect our outcomes of interest only

However, we follow Donaldson and Keniston (2016) in using levels of the climate variables without including their interactions.

through their effects on respiratory deaths in 1917-1920 (test for relevance), we report regressions in columns (4)-(6) of Table 3 that show that October's mean temperature and precipitation have no impact on respiratory death rates in the 1913-1915 time period.[22] Our conservative estimates are from the saturated specifications in columns (3) and (6) of Table 3. In the robustness section that follows, we show that our instruments have no direct impacts on the outcomes that we evaluate in 1920 and 1940, thus underlining that they satisfy the exclusion restriction.

October's temperature and rainfall are the instruments in all cases except when the outcome is agricultural productivity. This is because studies have documented the possible relationship between climatic conditions and agricultural output in the Age of Mass Migration (Hatton and Williamson, 1998, Solomou and Wu, 1999). Instead, we construct another instrument based on Acemoglu and Johnson (2007) and Percoco (2016). This instrument relies on the baseline regional distribution of deaths by respiratory infections in each district in $t < 1917$, and leverages the surge in influenza deaths during the pandemic years of 1917-1920. Let $AggFlu_t$ denote the aggregate number of deaths from influenza in Sao Paulo in year t. Let $\bar{s}_{d1915}$ be the proportion of deaths caused by influenza in a baseline period in district d. We choose the year 1915 as baseline as data are available for a representative number of districts in that year. We can reasonably assume that the baseline distribution of deaths by respiratory infections in each district and Sao Paulo's aggregate influenza mortality rate in a later year t are exogenous. So, as Percoco (2016) demonstrates, this instrument controls for any bias that results from the endogenous spatial distribution of the pandemic. For $t > 1915$, we impute the respiratory deaths as $(\bar{s}_{d1915} \times AggFlu_t)$. We standardize this value by pop_{dt}, the population of district d in year t. Equation 3 shows $IVResp_{dt}$, this constructed instrument for respiratory deaths:

$$IVResp_{dt} = \bar{s}_{d1915} \frac{AggFlu_t}{pop_{dt}} \tag{3}$$

Equation 3 exploits two sources of variation: the cross-sectional variation in the proportion of deaths caused by influenza for each district in 1915, and the time-series variation induced by changes in the aggregate mortality rate. The validity of the instrument rests on the assumption that the district-specific characteristics that determined $\bar{s}_{d1915}$ do not independently predict the future spatial distribution of the disease, that is, $\bar{s}_{d1915}$ should impact respiratory deaths only through its effect on $IVResp_{dt}$. This assumption may be violated however if the baseline proportion in 1915 is correlated with future changes in population, literacy rates, or income. Or

[22] As another falsification test, Table A5 in the Appendix reports the first stage regression results using April's temperature and rainfall, conditional on the same set of controls used for the preferred first stage regression on October's climate instruments. We also test for the relevance and strength of April's instruments holding October's temperature and precipitation constant. The results suggest that April's instruments are weak for both 1917-1920 and 1913-1915 as the F-statistic remains below 5. We obtain similar results when temperature and rainfall for May or June are used instead. Further, as noted above, Table A6 in the Appendix reports the first stage regressions with deviations of October's mean temperature and precipitation from their long-term averages over the 1901-2016 time period. As is clear, the F-statistic is below the required benchmark level.

the district-specific features that determined the baseline weight affect the evolution of social and economic conditions during the pandemic years. To net out such influences, we interact a district's pre-pandemic initial characteristics with year dummies. We do likewise for the geographic variables. Further, we use the per capita public expenditures on cleaning, waste disposal and maintenance of good sanitary conditions available for 132 municipalities in 1910 to ensure the assumption on which the validity of the instrument rests, holds. The first stage regression (as a function of the identifying instrument only) is shown in equation 4, with results reported in Table 4 :

$$Flu_{dt} = \gamma_0 + \gamma_1 IVResp_{dt} + \omega_{dt} \tag{4}$$

Table 4 indicates that this synthetic instrument is strong with F-statistics above the required threshold of 10. Morever, the coefficient on the constructed instrument $IVResp_{dt}$ is positive and significant, and robust to the inclusion of fixed-effects, and district-level geographical, health and initial controls. However, compared to the climatic instrumental variables discussed above, using the predicted influenza mortality rate in equation 3 reduces the sample size. Hence we use this constructed instrument only for agriculture-related outcomes.

Finally standard errors are clustered at the micro/mesoregion level in all our models.[23] The underlying intuition is that neighboring districts and municipalities share similar unobservable features, and most 1920 census districts belonged to larger stratifications sharing common initial conditions.[24]

6 Short-term Effects in 1920

6.1 Demographic Indicators

We begin our analysis by evaluating the pandemic's immediate impact on infant mortality, sex ratios at birth, and on the normalized number of still births. Panel A in Table 5 reports results for the infant mortality rate in 1920. Both OLS and 2SLS methodologies yield positive and significant coefficients in all specifications, with and without the inclusion of region fixed-effects and initial and region-level controls (including dummies for the presence of water and sewage systems, hospitals, nursing homes, specialized maternity hospitals, geographic controls, literacy rates, and other demographic controls). The IV estimates are larger than the OLS results indicating that the latter are downward biased, possibly due to attenuation bias in the presence of classical measurement error.[25] In the fully saturated model, all else equal, a unit increase in respiratory death

[23] We have information on 13 meso-regions and 57 micro-regions of Brazil in our sample.

[24] Bertrand et al. (2003) notes the importance of clustering standard errors at the largest sensible aggregation.

[25] Throughout our analysis, the IV coefficients are larger in magnitude than the OLS estimates. This indicates that attenuation bias from measurement error is substantial in our case. Note that the standard errors are also

rates increased infant mortality by 0.019. Given the mean value of infant mortality in the our sample, this is a 0.09% increase. Our estimates of the impact of the pandemic on still birth rates–loss of the baby before the 20th week of pregnancy in 1920–are presented in Part B of Table 5. The instrumental variable effects in Panel B indicate that, all else equal, there was a significant increase in still birth rates by 21 log points, a 0.41% increase.

Table 6 reports results for sex ratios at birth in 1920. Controlling for adult sex ratios in 1920, the share of people of different races, the initial normalized number of midwives, and the presence of water, sewage systems, hospitals, nursing homes and specialized maternity hospitals, as well as geographic controls, we find a negative impact on sex ratios as reported in Panel B columns (1) to (3). In keeping with the intuition that male fetuses are relatively more vulnerable to shocks as compared to female fetuses, the IV results indicate that sex-ratios declined (by about 40.74% relative to the mean), that is, relatively fewer males were born in the immediate aftermath of the pandemic.[26]

6.2 Literacy Rates

There is substantial evidence that health shocks can have long-lasting consequences on educational attainment (Currie and Stabile, 2006, Currie, 2009, Parman, 2013). We test this hypothesis by estimating the impact of influenza deaths on male and female literacy rates for different age groups. The 1920 census reports literacy figures for those who can read and write, but we do not know whether formal schooling took place.[27] Intuitively, we do not expect to find significant impacts for younger age groups in 1920, especially those who have yet to start schooling. But this might not be the case for older age-groups. Table 7 shows that this is indeed the case. The 2SLS results shown in Panel B of this table suggest that there is no effect on male and female literacy rates (of all age groups) but that the literacy rate for males aged 15 and above rose in the short-run whereas the coefficient for females while positive, is measured with error. This is similar to the impact on male literacy in Donaldson and Keniston (2016). An explanation for this may be the fact that since men are more likely to be in the labor market, and since relatively weaker prime-age men (possibly from the lower classes) died in greater numbers, the average literacy rate increased as a consequence of this type of differential selection.

larger for the IV coefficients relative to the OLS coefficients, as is usually true in 2SLS analyses.

[26] It is hard to reconcile the size of this coefficient with the impact on infant mortality rates as by definition, children have to be born to count as part of the sex ratio.

[27] The 1920 census provides the number of literate and illiterate individuals at the district level.

6.3 Agricultural Productivity

We focus on the agricultural sector and our outcomes of interest in 1920 are the output of coffee, rice and maize in tons, normalized by total population to obtain per capita values. The results are reported in Table 8. There are immediate sizeable effects on per capita volumes of coffee, rice, and maize in 1920, indicating that the pandemic caused a decrease in agricultural productivity.[28] The 2SLS estimate for per capita coffee production translates into a 21% decline relative to the mean while those for per capita rice production and per capita maize production represent a 47% and 25% decline relative to their respective means.[29]

7 Long-term Effects in 1940

7.1 Literacy Rates

As discussed above, we use the 1940 census to study impacts beyond the immediate short-run.[30] The Census of 1940 included 7 questionnaires related to literacy. Compared to 1920, however, questions were not limited only to the ability to read and write. We are therefore able to compute average municipality's literacy rate for people receiving some form of instruction. Given the evidence in Musacchio et al. (2014) that Sao Paulo experienced a rapid increase in the number of schools, students and teachers during the 1889-1930 period, we control for the share of total municipality expenditure allocated to education and health in 1940, the number of teachers (per person aged 10 and above), the share of total education expenditure spent on school supplies and on teachers' salaries. Other controls for initial conditions, share of people who were slaves, race measures and the urbanization rate as well as controls for altitude are also included in these regressions.

In Table 9 which reports a broad cohort style analysis, we find that female literacy was negatively impacted in districts with greater influenza exposure whereas male literacy rates were positively affected.[31] For the age group 20-29 years old, the cohort that was directly impacted by the 1918

[28] Table A7 in the Appendix shows that wages for coffee workers rose.

[29] In order to explain the impact on coffee in 1920, we collected data on coffee production in 1917/1918 and 1918/1919, and compared the means of these based on a binary variable that represented the intensity with which regions were affected by the pandemic (a binary variable that takes a value of 1 if the district reports respiratory death rates that were above the median death rate for all districts in the sample). We find that areas with above median respiratory death rates produced on average 17.5% more coffee in 1917/1918 and 11.7% more in 1918/1919. Thus relatively more coffee-producing regions were hardest hit by the pandemic. We are unable to do likewise for rice or maize. However, the fact that these crops are relatively more labor-intensive than coffee may explain the larger size of the coefficients in columns (2) and (3) in comparison to that in column (1).

[30] The 1940 census provides municipality-level data.

[31] We should be clear that we do not know whether these individuals were born in these districts or they were in the district when enumerated during the Census.

pandemic, we find that respiratory death rates from 1917-1920 lead to a significant fall in literacy rate for females. No such effect is found for female aged 30-39 years old who were already eight years old or older when the pandemic arrived. Alternatively, we find that there is a significant positive increase in male literacy rate for those aged 20-29 and 30-39 years old (the F-statistic for those in the 30-39 age group gives us little confidence in this result though). The 2SLS coefficient for 20-29 year old males indicates a 7.80% rise in literacy rates relative to the mean.

In Table 10, we further explore the pandemic's effects on literacy rates for broader age groups and on alternative measures of educational attainment. We estimate the impact on males and females aged 5 and above and aged 18 and above in columns (1) to (4) and again find a significant positive impact on the male literacy rate in column (1) and no effect on the female literacy rate in column (2) of Table 10. The estimated effects in columns ((3) and (4) for the older age group is similar. These results lend credence to the hypothesis that the 1918 pandemic had long-lasting repercussions on human capital.

We provide several explanations for the positive impact on male literacy in the aftermath of the pandemic. First, as noted above, excess male mortality in response to health shocks of this nature implies that the males that are born are positively selected. This could play a role in understanding the positive literacy coefficient. Further, investing in males/boys at the expense of females/girls is entirely consistent in poor country contexts where labor markets may be segmented for example. Second, positive impacts on men/boys have been found in other contexts. In the aftermath of the 2004 tsunami in Indonesia for example, Frankenberg et al. (2013) find that literacy rates for males rose. Other studies where male literacy has been found to have increased include Donaldson and Keniston (2016) and Parman (2013). Third, an increase in male literacy rates are also found when we apply the Almond (2006) baseline model to the Brazilian IPUMS data from 1960-1980. Finally, the education parameters that other researchers have considered and where negative impacts have been found are high school graduation rates and years of schooling. Although not inconsistent with literacy rates, these variables are different.

7.2 Hospital Admissions

We compute hospital in-patient admissions per 1000 inhabitants and the normalized number of people with disabilities admitted to nursing homes and asylums from the 1940 Census. The results are reported in Table 11. While we find no statistical significance for the nursing homes/asylum admissions, we find that respiratory deaths have a positive significant impact on the in-patient hospital admissions rates up to two decades after the pandemic's arrival in Sao Paulo. The 2SLS coefficient in column (1) indicates a 33.03% increase in this measure relative to the mean.

7.3 Agricultural Productivity

Table 12 reports the results for aggregate measures of productivity for the primary sector in 1940. We calculate the value of primary sector's output per employee (columns 1 and 2) and per establishment (columns 3 and 4), in logs. The results indicate that there are measurable drags on labor productivity from the pandemic up to twenty years after the event.[32] A one unit increase in respiratory deaths (per 1000) leads to a 0.45 log point decline in the value of the primary sector per employee. We obtain a coefficient of similar magnitude for the value per establishment. Relative to the averages for these measures of productivity, these impacts on productivity are significant.[33]

8 Robustness Checks

We conduct specification checks to verify the robustness of these results. As noted above, Table A5 in the Appendix reports the first stage regression results using April's temperature and rainfall, conditional on the same set of controls used for the preferred first stage regression on October's climate instruments. The results suggest that April's instruments are weak for both 1917-1920 and 1913-1915 as the F-statistic remains below 5. We obtain similar results when temperature and rainfall for May or June are used instead. Further, Table A6 in the Appendix reports the first stage regressions with deviations of October's mean temperature and precipitation from their long-term averages over the 1901-2016 time period. As is clear, the F-statistic is below the required benchmark level. Next, we test to ensure that the sample of births in the flu-years are representative. We rely on male and female literacy rates in 1872 as a measure of parental characteristics, and compare these to the rates in 1920 in order to rule out selection on parental observables. The test of differences in means between the 1872 and 1920 literacy rates cannot reject that these are the same ($p = 0.41$). We then generate an indicator variable that takes a value of 1 if a district in 1918 had flu exposure above the median value in the sample (that is, if respiratory deaths per 1000 inhabitants exceeded 1.21 in 1918). We then test for differences in sample means in literacy rates across these two groups. These results are reported in Table 13. Again, test results indicate that the samples are comparable along this dimension.[34]

[32] We added additional controls including a proxy for mechanization and capital utilization–the number of plows used per establishment, disaggregated forms of expenditure per agricultural establishment on new seeds, fertilizers and insecticides, on salaries, and on the acquisition of new machinery and animals in 1939.

[33] Table A8 in the Appendix reports different measures of productivity in 1940 for small-scale and large-scale businesses in farming, agriculture, and livestock. The results indicate that by 1940, there had been declines in land productivity as measured by the value of the agricultural sector per hectare of cultivated land. The disaggregated analysis indicates that the negative effect on labor productivity is stronger for the agricultural sector relative to livestock and farming.

[34] Given data constraints, we have no other variable to benchmark parental characteristics.

In Table 14, we carry out falsification tests by demonstrating that the respiratory death rate during a non-pandemic year (1915) has no impact on outcome variables either in 1920 or in 1940. All 2SLS (and OLS) coefficients are statistically zero for sex ratios and literacy rates in 1920, and literacy rates in 1940. Further, since our empirical strategy relies on the predicted mortality instrument, we show in Table 15 that the baseline weight ($\bar{s}_{d1915}$), the share of total deaths caused by influenza-related causes in 1915, has no direct predictive power for economic variables before the 1918 pandemic or in the years we consider. Since we do not have data on a lot of variables between 1915 and 1920, we use still births in 1915 and wages in 1915. The last two columns of Table 15 provide further supporting evidence by demonstrating that the constructed weight has no direct impact on productivity outcomes in 1920 or 1940. This strengthens the case that the constructed cross-sectional baseline weight acts only through its effect on respiratory death rates during the pandemic years.

Finally, in order to ensure that the instruments have no direct effects on the outcomes we study, we regress the instruments - October's mean temperature and precipitation - on outcomes from 1920 and 1940. These results are reported in Table 16. Net of fixed effects, region-level and initial period controls, the estimates in Table 16 indicate that the instruments have no direct impacts.[35]

9 Conclusion

We use data from the 1872, 1920, and 1940 Brazilian censuses for more than 300 geographical units in the State of Sao Paulo, Brazil, to study the impact of the 1918 influenza pandemic across different time horizons. We match these data to district-level deaths by cause during the pandemic years (1917-1920), and combine this matched dataset with an array of historical and geographical information drawn from several archival documents. We show that the pandemic had significant short-run demographic, literacy and productivity impacts in 1920. We then study dynamics on literacy rates, in-patient hospital admissions, and productivity in 1940, to demonstrate lasting effects. These results are robust to a variety of falsification and specification checks. Our results underline that historical determinants of development can have persistent effects on outcomes, and that a comprehensive understanding of the legacy of health shocks provides scope for targeted present-day policy-making in order to ameliorate some of the negative consequences.

[35] Results for the full set of outcomes are available on request.

References

Acemoglu, D. and S. Johnson (2007). Disease and development: The effect of life expectancy on economic growth. *Journal of Political Economy 115*(6), 925–985.

Acemoglu, D., S. Johnson, and J. Robinson (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review 91*(5), 1369–1401.

Almond, D. (2006). Is the 1918 influenza pandemic over? long-term effects of in utero influenza exposure in the post-1940 us population. *Journal of Political Economy 114*(4), 672–712.

Almond, D. and B. Mazumder (2005). The 1918 influenza pandemic and subsequent health outcomes: An analysis of sipp data. *American Economic Review 95*(2), 258–262.

Alonso, W., F. Nascimento, R. Acuna-Soto, C. Schuck-Paim, and M. Miller (2016). The 1918 influenza pandemic in florianopolis: A subtropical city in brazil. *Vaccine 29*, B16–B20.

Alsan, M. (2015). The effect of the tsetse fly on african development. *American Economic Review 105*(1), 382–410.

Bakken, M. and S. Husoy (2016). The long term impact of the 1918 influenza pandemic in norway. *Master's Thesis*.

Barata, R. (2000). Cem anos de endemias e epidemias. *Ciencia & Saude Coletiva 5*(2), 333–345.

Barker, D. (1995). Fetal origins of coronary heart disease. *BMJ 311*, 171–4.

Bassanezi, M. (2013). Uma tragica primavera. a epidemia de gripe de 1918 no estado de sao paulo, brasil. *ESTADO DE SAO P*, 73.

Beach, B. and W. Hanlon (2017). Coal smoke and mortality in an early industrial economy. *The Economic Journal 128*(615), 2652–2675.

Beach, B., J.Ferrie, and M.Saavedra (2018). Fetal shock or selection? the 1918 influenza pandemic and human capital development. *National Bureau of Economic Research.* (No. w24725).

Bertolli, F. (2003). A gripe espanhola em sao paulo, 1918: epidemia e sociedade, sao paulo, paz e terra. *Colecao Sao Paulo 5*.

Bertrand, M., E. Duflo, and S. Mullainathan (2003). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*, 249–275.

Bertucci-Martins, L. (2005). Entre doutores e para os leigos: Fragmentos do discurso medico n influenza de 1918. *Historia, Ciencias, Saude-Manguinhos 12*(1).

Bleakley, H. (2010). Malaria eradication in the americas: A retrospective analysis of childhood exposure. *American Economic Journal: Applied Economics 2*(2), 1–45.

Caruso, G. (2015). Intergenerational transmission of shocks in early life: Evidence from the tanzania great flood of 1993. *SSRN 2560876*.

Caruso, G. and S. Miller (2015). Long run effects and intergenerational transmission of natural disasters: A case study of the 1970 ancash earthquake. *Journal of Development Economics 117*, 134–150.

Chul Hong, S. and Y. Yun (2017). Fetal exposure to the 1918 influenza pandemic in colonial korea and human capital development. *Seoul Journal of Economics 30*(4), 353–383.

Clay, K., J. Lewis, and E. Severnini (2018). Pollution, infectious disease, and mortality: Evidence from the 1918 spanish influenza pandemic. *The Journal of Economic History 78*(4), 1179–1209.

Clay, K., J. Lewis, and E. Severnini (2019). What explains cross-city variation in mortality during the 1918 influenza pandemic? evidence from 440 u.s. cities. *Economics & Human Biology 35*, 42–50.

Cook, C., J. Fletcher, and A. Forgues (2018). Multigenerational effects of early life health shocks. *National Bureau of Economic Research* (No. w25377).

Crosby, A. (1989). America's forgotten pandemic: The influenza of 1918.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature 41*(1), 87–122.

Currie, J. and M. Stabile (2006). Child mental health and human capital accumulation: The case of adhd. *Journal of Health Economics 25*(6), 1094–1118.

Dell, M. (2010). The persistent effects of peru's mining mita. *Econometrica 78*(6), 1863–1903.

Donaldson, D. and D. Keniston (2016). Dynamics of a malthusian economy: India in the aftermath of the 1918 influenza. *Unpublished Manuscript.*.

Frankenberg, E., B. Sikoki, C. Sumantri, W. Suriastini, and D. Thomas (2013). Education, vulnerability, and resilience after a natural disaster. *Ecology and Society 18*(2), 16.

Garrett, T. (2007). Economic effects of the 1918 influenza pandemic.

Glaeser, E., R. L. Porta, F. L. de Silanes, and A. Shleifer (2004). Do institutions cause growth? *Journal of Economic Growth 9*(3), 271–303.

Guiso, L., P. Sapienza, and L. Zingales (2016). Long-term persistence. *Journal of the European Economic Association 14*(6), 1401–1436.

Hamoudi, A. and J. Nobles (2014). Do daughters really cause divorce? stress, pregnancy, and family composition. *Demography 51*(4), 1423–1449.

Hanlon, W. (2019). Coal smoke, city growth, and the costs of the industrial revolution. *The Economic Journal*, forthcoming.

Hatton, T. and J. Williamson (1998). The age of mass migration: Causes and economic impact. *Oxford University Press*.

Hochman, G. (2016). The sanitation of brazil: Nation, state, and public health, 1889-1930. *University of Illinois Press*.

Huillery, E. (2009). History matters: The long-term impact of colonial public investments in french west africa. *American Economic Journal: Applied Economics 2*(1), 176–215.

Karlsson, M., T. Nilsson, and S. Pichler (2014). The impact of the 1918 spanish flu epidemic on economic performance in sweden: An investigation into the consequences of an extraordinary mortality shock. *Journal of Health Economics 36*, 1–19.

Kobasa, D., A. Takada, K. Shinya, M. Hatta, P. Halfmann, S. Theriault, H. Suzuki, H. Nishimura, K. Mitamura, N. Sugaya, and T. Usui (2004). Enhanced virulence of influenza a viruses with the haemagglutinin of the 1918 pandemic virus. *Nature 431*(7009), 703.

Lin, M. and E. Liu (2014). Does in utero exposure to illness matter? the 1918 influenza epidemic in taiwan as a natural experiment. *Journal of Health Economics 37*, 152–163.

Loo, Y. and M. Gale (2007). Influenza: Fatal immunity and the 1918 virus. *Nature 445*(7125), 267.

Massad, E., M. Burattini, F. Coutinho, and L. Lopez (2007). The 1918 influenza a epidemic in the city of sao paulo, brazil. *Medical Hypothesis 68*(2), 442–445.

Musacchio, A., A. Fritscher, and M. Viarengo (2014). Colonial institutions, trade shocks, and the diffusion of elementary education in brazil, 1889-1930. *Journal of Economic History 74*(3), 730–766.

Nelson, R. (2010). Testing the fetal origins hypothesis in a developing country: Evidence from the 1918 influenza pandemic. *Health Economics 19*(10), 1181–1192.

Neumayer, E. and T. Plumper (2007). The gendered nature of natural disasters: The impact of catastrophic events on the gender gap in life expectancy. *Annals of the Association of American Geographers 97*(3), 551–566.

Noymer, A. and M. Garenne (2000). The 1918 influenza epidemic's effects on sex differentials in mortality in the united states. *Population and Development Review 26*(3), 565–581.

Nunn, N. and L. Wantchekon (2011). The slave trade and the origins of mistrust in africa. *American Economic Review 101*(7), 3221–3252.

Parman, J. (2013). Childhood health and sibling outcomes: The shared burden and benefit of the 1918 influenza pandemic. *NBER Working Paper No. w19505*.

Percoco, M. (2016). Health shocks and human capital accumulation: The case of the spanish flu in italian regions. *Regional Studies 50*(9), 1496–1508.

Polozov, I., L. Bezrukov, K. Gawrisch, and J. Zimmerberg (2008). Progressive ordering with decreasing temperature of the phospholipids of influenza virus. *Nature Chemical Biology 4*(4), 248.

Putterman, L. and D. Weil (2010). Post-1500 population flows and the long-run determinants of economic growth and inequality. *The Quarterly Journal of Economics 125*(4), 1627–1682.

Richter, A. and P. Robling (2013). Multigenerational effects of the 1918-19 influenza pandemic in sweden. *Swedish Institute for Social Research 5*.

Rocha, R., C. Ferraz, and R. Soares (2017). Human capital persistence and development. *American Economic Journal: Applied Economics 81*(1), 51–57.

Sanders, N. and C. Stoecker (2015). Where have all the young men gone? using sex ratios to measure fetal death rates. *Journal of Health Economics 41*, 30–45.

Slutsky, D. and R. Zeckhauser (2018). Sunlight and protection against influenza. *National Bureau of Economic Research No. w24340*.

Solomou, S. and W. Wu (1999). Weather effects on european agricultural output, 1850-1913. *European Review of Economic History 3*(3), 351–373.

Tamerius, J., J. Shaman, W. Alonso., K. Bloom-Feshbach, C. Uejio, A. Comrie, and C. Viboud (2013). Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS Pathogens 9*(3), e1003194.

Figure 1: A Newspaper Extract in October 1918

GAZETA DE NOTICIAS

Anonyma GAZETA DE NOTICIAS.

A GRANDE EPIDEMIA

As providencias do governo de muito pouco valeram até agora

Continuamos entregues á Divina Providencia : E os casos fataes augmentam :

A gravidade da situação e as providencias do governo

Os mortos de hontem

A Caixa Economica

O seu fechamento creou uma situação acti-ctiva para o publico

Providencias do governo

O expediente nas delegacias de saude

O serviço sanitario na Guanabara

O aviamento de receitas

Uma nota da Directoria de Saude Publica

O trabalho no palacio do Cattete

A Gazeta de Notícias critica o governo federal pelo imobilismo diante da disseminação da epidemia (imagem: Biblioteca Nacional)

Notes: This extract conveys that the government was criticized for its lack of involvement during the pandemic in 1918. Translated headline: the government's input of very little value until now. Source: Biblioteca Nacional.

Figure 2: Headline of the Correio da Manha in October 1918

SÃO INNUMEROS OS PEDIDOS DE SOCCORROS AO PALACIO DO CATTETE

O governo faz um appello aos medicos, pharmaceuticos e estudantes, para auxiliarem a Saude Publica nesta difficil emergencia

Na guarda-moria da Alfandega

O governo determinou hontem novas providencias

O caso dos coveiros

O que fez officialmente o Commissariado

O presidente da Republica fala com o sr. Bulhões

A Telephonica soffre o primeiro golpe

A deserção na Saude Publica

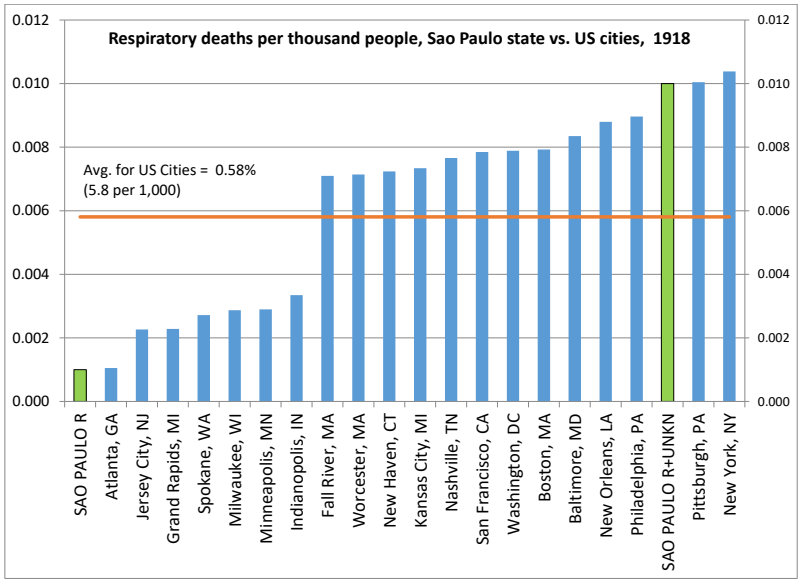
Em Nictheroy

Em Petropolis

VARIAS

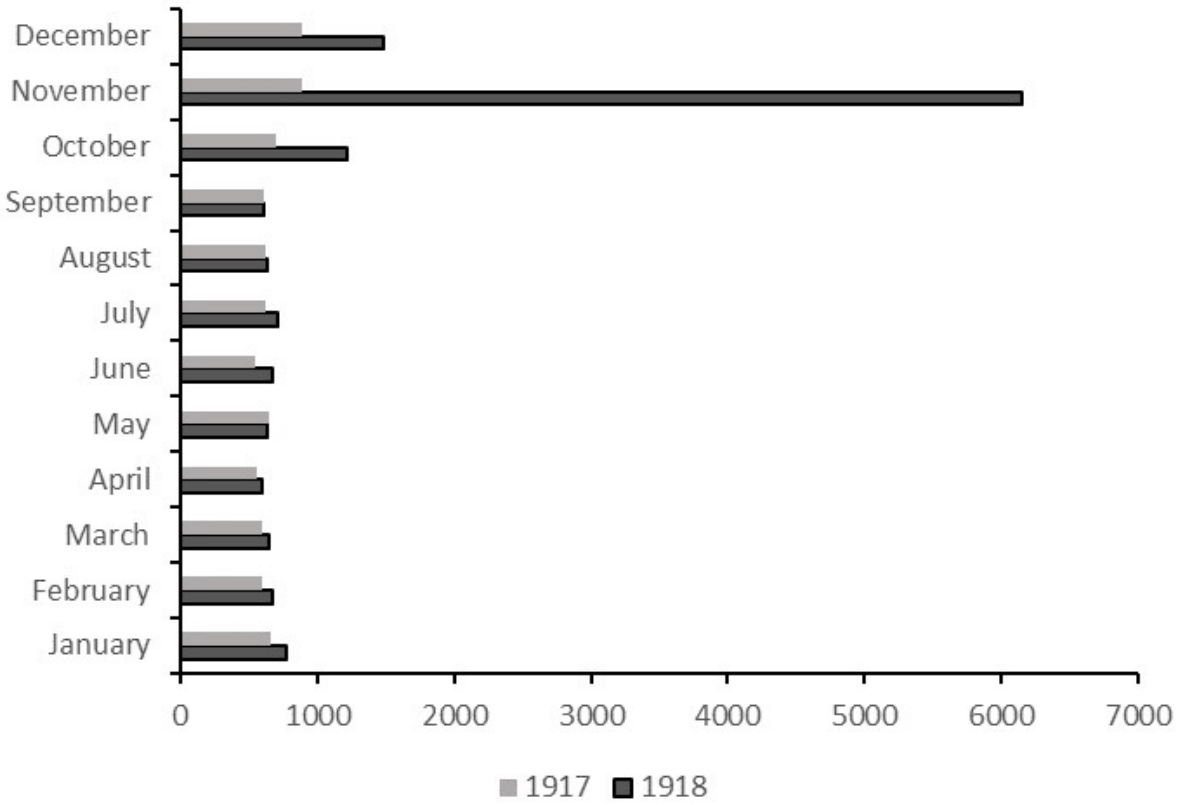
Notes: This extract conveys that the government needed help from medical experts and students at the peak of the Spanish Flu outbreak in October 1918. Translated headline: The government appeals to doctors, chemists and students to help public health during this period of emergency. Source: Collections of the Morning Mail in the Brazilian Digital Newspaper Archive.

Figure 3: Respiratory Death Rates for US Cities and Sao Paulo (State)



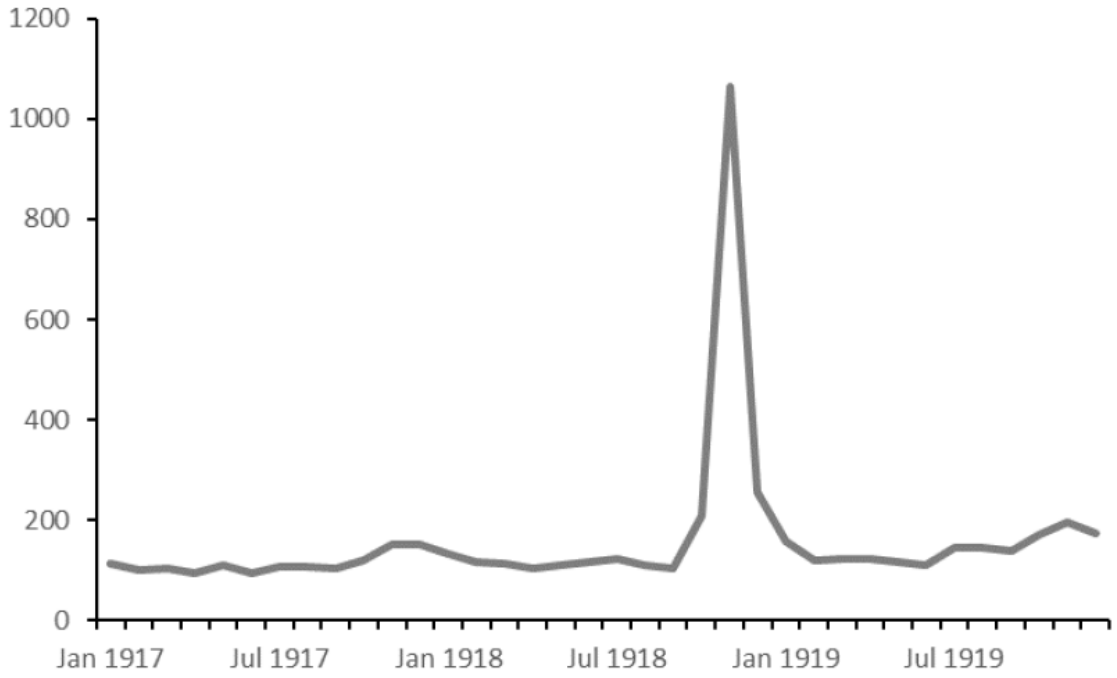
Notes: Source for US cities data is Crosby (1989). The green bar on the left shows the influenza-related death rates proxied by respiratory deaths only while the green bar on the right includes both deaths caused by respiratory and unknown causes.

Figure 4: Monthly Deaths for 1917 and 1918 for Sao Paulo (City)



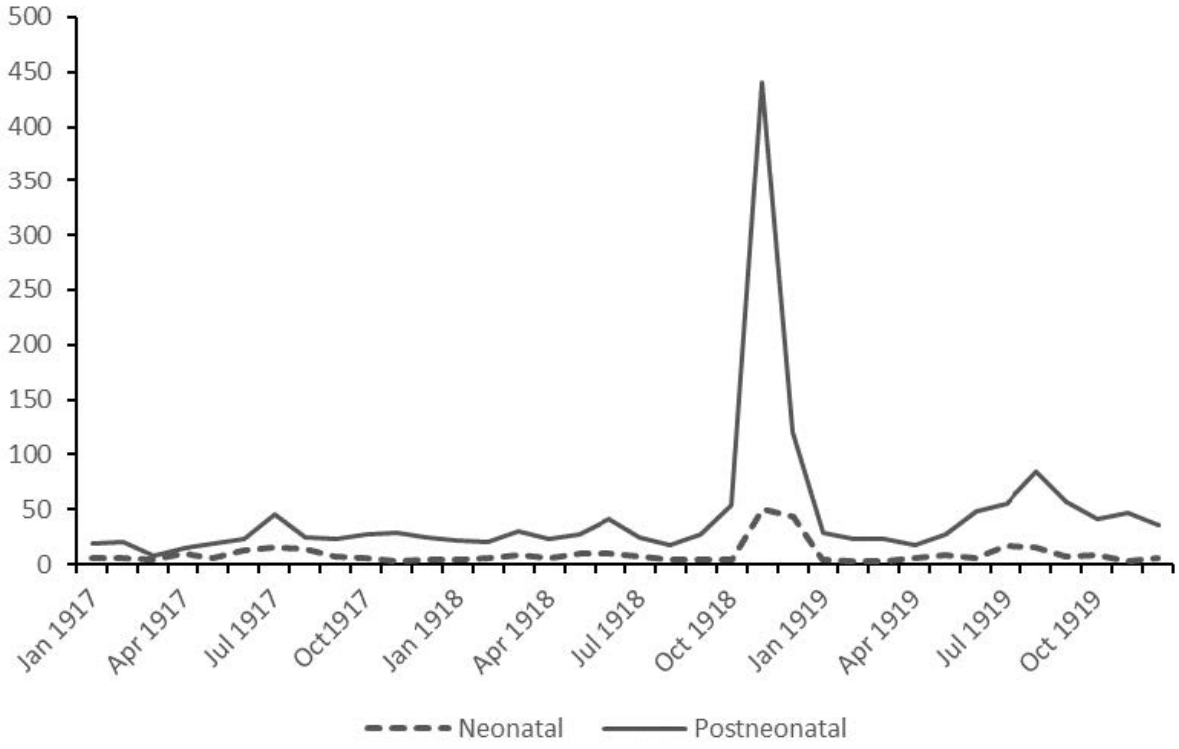
Source: Annual Statistical Reports Vol. 1 (1917, 1918, 1919) for the Sao Paulo (City)

Figure 5: Monthly Total Deaths per 100,000 of the Population (January 1917-December 1919)



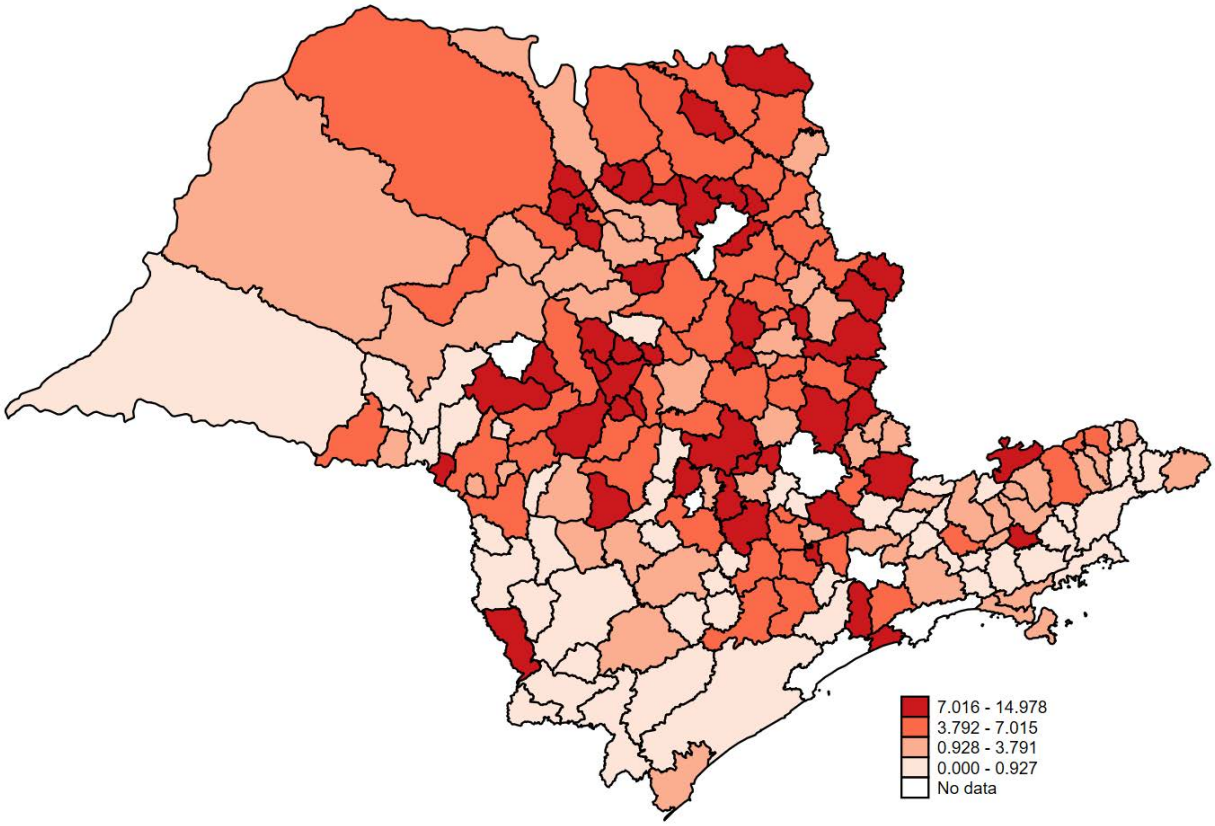
Source: Authors' calculations from the Annual Statistical Reports Vol. 1 (1917, 1918, 1919) for the Municipality Capital (Sao Paulo City)

Figure 6: Monthly Neo-Natal and Post-NeoNatal Influenza-Related Mortality Rate (Deaths per 1000 Live Births for January 1917-December 1919)



Source: Authors' calculations from the Annual Statistical Reports Vol. 1 (1917, 1918, 1919) for the Municipality Capital (Sao Paulo City)

Figure 7: The Spatial Variation in Influenza-Related Deaths per 1000 of the Population in Sao Paulo (State), 1917-1920



Source: Authors' calculations from the Annual Statistical Reports Vol.1 (1917, 1918, 1919, 1920). The data is aggregated up to the municipality level for each year during this period. The 1920 boundaries are used to consider the variation.

Figure 8: The Spatial Variation in Respiratory and Unknown Deaths per 1000 of the Population in Sao Paulo (State), 1918

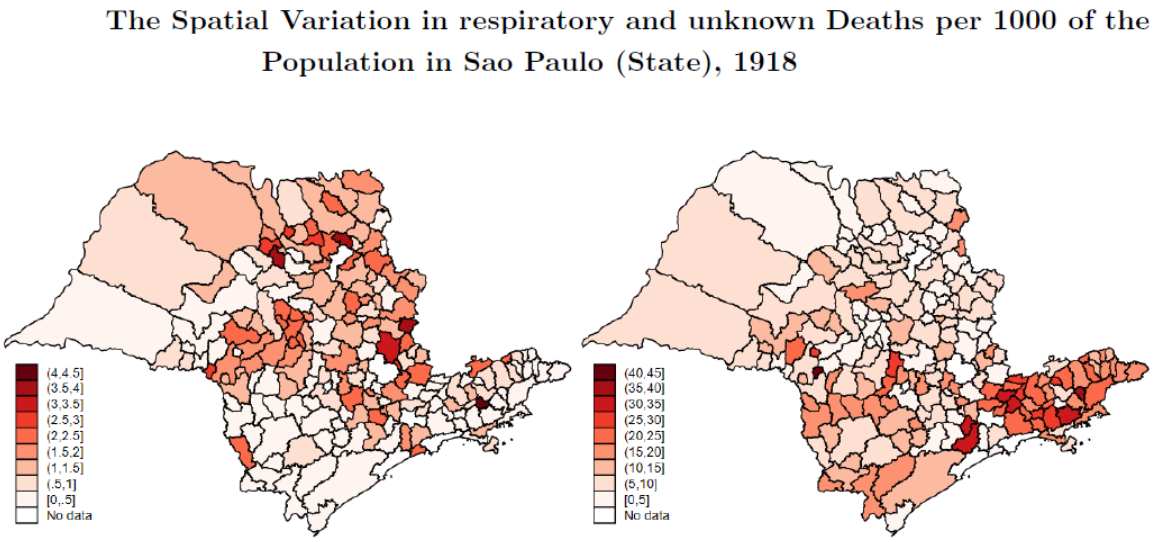
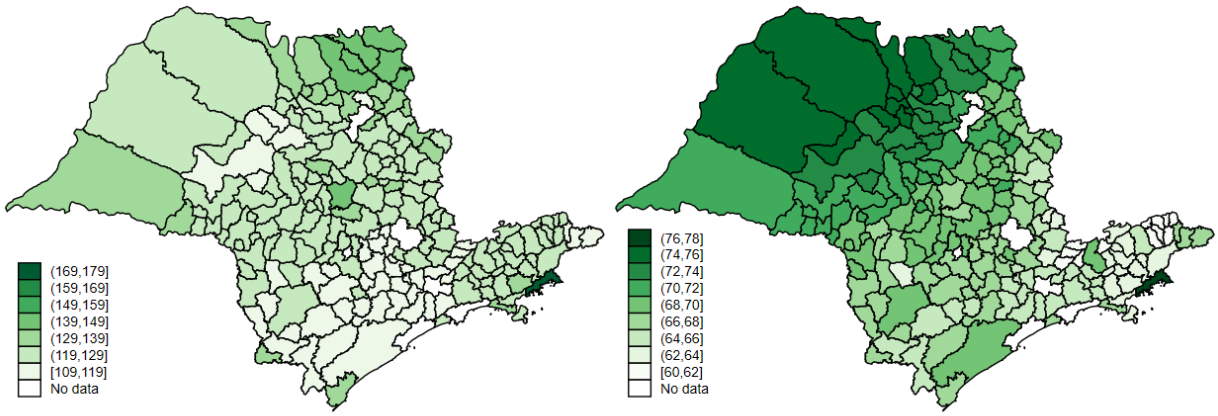


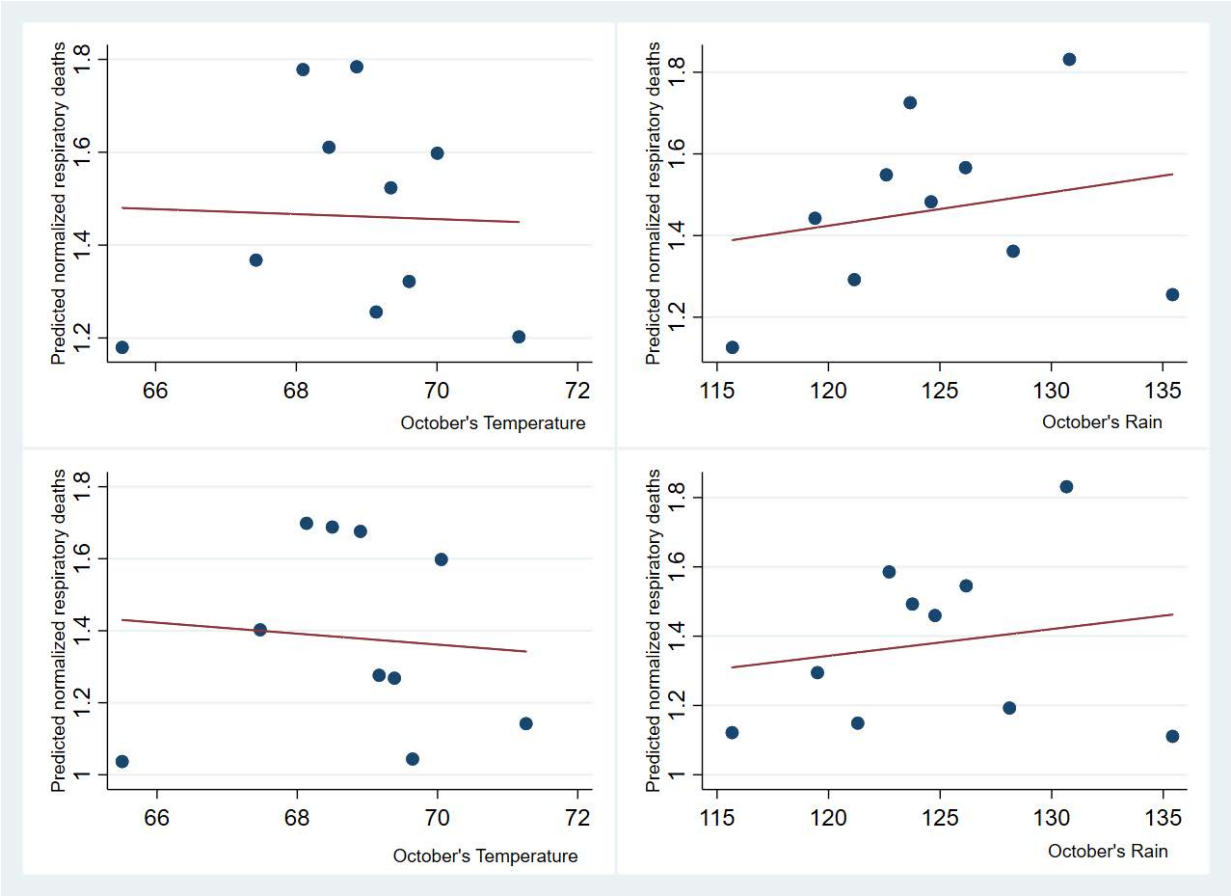
Figure 9: Average October's Rain and Temperature, 1901-1930



(a) Average Precipitation (mm) in October

(b) Average Temperature (Fahrenheit) in October

Figure 10: Binscatter Plots for Respiratory Death Rates and October's Rain and Temperature



Notes: Panel A (top: 1917-1920) and Panel B (Bottom: 1918/1919 only). Regressions condition on geographical controls that include altitude, longitude and latitude.

Table 1: Summary Statistics for Dependent Variables

	Mean (1)	Standard Dev (2)
Panel A: Variables in 1920		
Infant mortality (per live births)	0.02	0.03
Still births (per total births)	0.05	0.03
Literacy rate (total)	0.23	0.08
Literacy rate (male)	0.28	0.09
Literacy rate (female)	0.18	0.08
Sex ratio at birth (Male babies/Female babies)	1.21	0.82
Coffee production per capita (tons)	0.31	0.45
Coffee production per establishment (tons)	14.88	14.16
Rice production per capita (tons)	0.29	0.59
Rice production per establishment (tons)	5.58	8.06
Maize production per capita (tons)	1.19	2.13
Maize production per establishment (tons)	19.67	12.68
Panel B: Variables in 1940		
Share receiving instruction and literate:		
Male (age 7-14)	0.82	0.089
Female (age 7-14)	0.84	0.10
Share receiving instruction and literate:		
Male (age 5-39)	0.82	0.09
Female (age 5-39)	0.85	0.35
Literacy rate:		
Male (10-19 years)	0.52	0.14
Female (10-19 years)	0.48	0.15
Male (20-29 years)	0.59	0.13
Female (20-29 years)	0.41	0.14
Male (30-39 years)	0.58	0.13
Female (30-39 years)	0.32	0.12
Male (Age 5 and above)	0.49	0.12
Female (Age 5 and above)	0.35	0.12
Inpatient Hospital Admissions (per 1000)	3.79	15.19
Inpatient Asylum Admissions (per 1000)	6.18	14.48
Value of primary sector per emp. (cr$1000)	1009.03	490.85
Value of primary sector per est. (cr$1000)	9470.19	7815.85

Notes: See Appendix 1 for data description and sources. Authors' calculations.

Table 2: Summary Statistics for Control Variables

	Mean (1)	Standard Dev (2)
Panel A: Deaths by Cause (Per 1000)		
Respiratory	1.74	2.03
Central Nervous System	0.74	0.96
Skin-related	0.14	0.17
Puerperal Sepsis	0.28	0.26
Circulatory system	1.37	5.68
Digestive	3.57	4.37
Panel B: Geography		
Altitude (in meters)	615.56	192.02
Latitude	-22.44	1.09
Longitude	-47.75	1.31
Distance to Capital (in Km)	236.34	113.60
October's Temperature (Fahrenheit)	68.43	3.76
October's Precipitation (mm)	124.15	8.39
Panel C: Variables from 1872 Census		
Literacy rate	0.18	0.09
Share of foreigners	0.03	0.03
Share of slaves	0.18	0.10
Share of race Branca	0.53	0.11
Share of race Parda	0.24	0.06
Share of race Preta	0.19	0.08
Share emp. in agriculture	0.64	0.11
Share emp. in manufacturing	0.13	0.05
Share with physical/mental diseases	11.56	6.02
Doctors (per 1000)	0.26	0.36
Chemists (per 1000)	0.18	0.22
Midwives (per 1000)	0.35	0.50
Population density (per sq.Km)	12.55	8.80
Panel D: Other Controls		
Health & sanitation exp. per capita (1910)	199.73	422.09
Dummy for frost intensity (1919)	0.71	0.45
Railway station dummy (1920)	0.70	0.46
Water & sewerage system dummy (1920)	0.20	0.40
No. hospitals&nursing homes (1920)	1.67	1.03
Taxes out of revenue (%,1920)	0.34	0.16
Sex ratio of adults (1920)	1.12	0.59
Share of foreigners (1920)	0.13	0.10
Share of male foreigners (1940)	0.06	0.04
Share of female foreigners (1940)	0.06	0.05
Share of exp. on health&education (1940)	0.61	0.12
Share of education exp. to school supplies (1940)	0.14	0.20
Share of education exp. to payment of teachers (1940)	0.81	0.49
Primary sector exp per est. (cr$1000,1940)	7.91	13.36
Agriculture exp per est.(cr$1000,1940)	5.82	7.06
Farming exp per est.(cr$1000,1940)	7.75	8.19
Livestock exp per est. (cr$1000,1940)	11.75	12.23
Number of plows per est.(1940)	1.54	0.63
Share of est. exp on salaries(1940)	0.70	0.41
Share of est. exp on new machinery(1940)	0.04	0.03
Share of est. exp on new animals(1940)	0.08	0.10
Share of est. exp on fertilizers(1940)	0.10	0.09

Notes: See Appendix 1 for data description and sources. Authors' calculations.

Table 3: First Stage Regressions: Dependent Variable is Respiratory Deaths (per 1000 of the Population)

	For 1917-1920			For 1913-1915		
	(1)	(2)	(3)	(4)	(5)	(6)
October's Mean Temperature	-0.245***	-0.246***	-0.252***	-0.320	-0.290	0.0337
	(0.056)	(0.056)	(0.056)	(0.361)	(0.464)	(0.338)
October's Mean Precipitation	0.032**	0.032**	0.029**	0.167	-0.001	-0.077
	(0.013)	(0.013)	(0.013)	(0.128)	(0.109)	(0.0394)
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Initial Controls of 1872	No	Yes	Yes	No	Yes	Yes
Geographical Control	No	No	Yes	No	No	Yes
F-Statistic	10.24	10.24	10.53	1.08	2.69	3.135
p-value for the F-Stat	[0.000]	[0.000]	[0.000]	[0.342]	[0.070]	[0.045]
Observations	939	937	933	414	413	410
R-squared	0.442	0.442	0.443	0.736	0.735	0.735

Notes: The table reports OLS regressions with clustered standard errors. In Column (1), we consider both instruments with only municipality fixed effects. In Columns (2) and (3), the controls are added sequentially. These include altitude, population density, share of employment in agriculture, manufacturing and in services/retail, the share of people who were slaves, the share of white people, the share of foreigners, of literate people, the normalized number of people with a mental and physical disease, the number of doctors, chemists and midwives (per 1000 people). p-values for the F-statistics are reported for the identifying instruments and are in square brackets. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$.

Table 4: First Stage Regressions for the Constructed Instrument : Dependent Variable is Respiratory Deaths (per 1000 of the Population)

	(1)	(2)	(3)	(4)
IVResp	0.237***	0.230***	0.223***	0.220***
	(0.043)	(0.051)	(0.049)	(0.069)
Region FE	No	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes
District-level controls	No	Yes	Yes	Yes
Initial Conditions of 1872	No	No	Yes	Yes
Initial Conditions x Year FE	No	No	Yes	Yes
1910 Heath and Sanitation Exp.	No	No	No	Yes
F-Statistic	30.22	20.81	21.01	10.20
p-value for the F-Statistic	[0.000]	[0.000]	[0.000]	[0.000]
Observations	853	544	544	353
R-squared	0.053	0.190	0.245	0.367

Notes: The table reports OLS regressions for the synthetic instrument (representing the interaction between the cross-sectional variation in the proportion of deaths caused by influenza in each district in 1915 and the time-series variation induced by changes in the aggregate mortality rate). Standard errors reported in parentheses are clustered at the district level. We use the variation for the period 1917-1920 only. In column (2), we add the year fixed effects, the district level controls: altitude, latitude, longitude, the distance to the capital, a dummy for railway station, the interaction between distance and the railway dummy. In column (3), the initial conditions of 1872 include the number of doctors, chemists and midwives (all refer to the numbers in the given occupation per 1000 of the population); the normalized number of people with a mental and physical disease, the literacy rate and the share of foreigners, and we add the interaction between the initial conditions and the year FE. Finally, in column (4), we add the 1910 health and sanitation expenditure, also interacted with the years. p-values for the F-statistics are reported for the identifying instruments and are in square brackets. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$

Table 5: Short-Run Effects on Infant Mortality and Still Births in 1920

	(1)	(2)	(3)	(4)
Part A: Dependent Variable: Infant Mortality				
OLS				
Norm. Respiratory Deaths	0.009***	0.008***	0.001***	0.009***
	(0.000)	(0.000)	(0.001)	(0.001)
2SLS				
Norm. Respiratory Deaths	0.017***	0.018***	0.020***	0.019***
	(0.003)	(0.005)	(0.006)	(0.007)
Region FEs	No	Yes	Yes	Yes
Region-level controls	No	No	Yes	Yes
Initial period controls from 1872	No	No	No	Yes
Mean for infant mortality	23.31	23.31	23.31	23.31
K-P F-stat (for infant mortality)	51.33	30.39	28.07	21.14
p-value for the F-stat (for infant mortality)	[0.000]	[0.000]	[0.000]	[0.001]
Observations	443	443	305	305
Part B: Dependent Variable: Log Still Births				
OLS				
Norm. Respiratory Deaths	-0.001	0.028	0.063*	0.061**
	(0.023)	(0.019)	(0.029)	(0.029)
2SLS				
Norm. Respiratory Deaths	0.038	0.199	0.193*	0.214**
	(0.123)	(0.130)	(0.102)	(0.084)
Region FEs	No	Yes	Yes	Yes
Region-level controls	No	No	Yes	Yes
Initial period controls from 1872	No	No	No	Yes
K-P F-stat (for log still births)	34.16	20.19	24.36	51.97
p-value for the F-stat (for log still births)	[0.000]	[0.001]	[0.000]	[0.001]
Mean for log still births	3.84	3.84	3.84	3.84
Observations	603	603	399	399

Notes: The table shows the OLS results (Panel A) and 2SLS results (in Panel B with IVResp as instrument). Standard errors reported in parentheses are clustered at the mesoregion level. We use the variation for the period 1917-1919 to evaluate the effects on 1920 infant mortality and still biths. In Column (1), we consider the regression without fixed-effects, controls and the initial controls of 1872. In Column (2), we add the region fixed-effects. In Column (3), the controls include a dummy for the presence of water and sewerage systems and one for the presence of hospitals, nursing homes and specialized maternity hospitals; the literacy rates for male and female aged 15 and above, region-level controls including altitude, latitude, longitude, distance to the capital city, a dummy for the presence of a railway line, and the interaction between distance and the railway dummy. In Column (4), we add the initial controls of 1872 and they include the number of doctors, chemists and midwives (all refer to the numbers in the given occupation per 1000 of the population), the normalized number of people with a mental and physical disease, share of people who were slaves, the share of employment in agriculture, manufacturing, and in services/retail, population density. The K-P F-stat refers to the Kleibergen-Paap F-statistic for weak instruments. ***$p < 0.01$, **$p < 0.05$ and *$p < 0.10$.

Table 6: The Short-run Effects on Sex Ratios at Birth in 1920

	Dependent Variable: Sex Ratio At Birth		
	(1)	(2)	(3)
Panel A: OLS			
Norm. Respiratory Deaths	-0.054**	-0.048*	-0.044
	(0.023)	(0.023)	(0.026)
Panel B: 2SLS			
Norm. Respiratory Deaths	-0.479**	-0.470***	-0.497***
	(0.221)	(0.143)	(0.173)
Region-level controls	No	No	Yes
Initial period controls from 1872	No	Yes	Yes
Geographical controls	No	No	Yes
K-P F-stat	7.98	10.24	19.60
p-value for the F-Stat	[0.007]	[0.003]	[0.000]
Mean of Dependent Variable	1.22	1.22	1.22
Observations	580	580	576

Notes: The table shows the OLS results (Panel A) and 2SLS results (in Panel B with October's rain and precipitation used as instruments). Standard errors reported in parentheses are clustered at the mesoregion level. We use the variation for the period 1917-1920 to evaluate the effects on 1920 sex ratios. In columns (1) we consider the regressions without controls and fixed-effects. In columns (2) and (3), we also include region-level controls (altitude, the sex ratio in 1920 for adults, a dummy for the presence of water and sewerage systems and one for the presence of hospitals, nursing homes and specialized maternity hospitals), and the initial controls include the normalized number of midwives, the shares of people of race Branca, Parda, Preta and Cabocla in 1872, the literacy rate. The K-P F-stat refers to the Kleibergen-Paap F-statistic for weak instruments. ***$p < 0.01$, **$p < 0.05$ and *$p < 0.10$.

Table 7: Literacy Rates by Gender and Age Groups in 1920

	All Ages			***Ages 15 and Above***		
	Male	Female	Total	Male	Female	Total
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: OLS						
Norm. Respiratory Deaths	0.010***	0.011***	0.011***	0.012***	0.013***	0.013***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)
Panel B: 2SLS						
Norm. Respiratory Deaths	0.021	-0.025	0.001	0.097**	0.009	0.064**
	(0.025)	(0.021)	(0.022)	(0.039)	(0.023)	(0.031)
Region-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Initial period controls from 1872	Yes	Yes	Yes	Yes	Yes	Yes
Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes
K-P F-stat	4.64	4.64	4.64	4.64	4.64	4.64
p-value for the F-Stat	[0.032]	[0.032]	[0.032]	[0.032]	[0.032]	[0.032]
Mean of Dependent Variable	0.281	0.164	0.226	0.418	0.221	0.326
Observations	669	669	669	669	669	669

Notes: The table reports the OLS results in Panel A and the 2SLS results in Panel B. The instruments used are October's rain and October's precipitation. Standard errors reported in parentheses are clustered at the mesoregion level. We use observations for the period 1917-1919. The initial controls of 1872 include altitude, the normalized number of people with a mental and physical disease, the literacy rate, the share of foreigners, the share of workers in agriculture, manufacturing and services, the share of people who were slaves, the share of people of race "Branca" (white), population density. The K-P F-stat refers to the Kleibergen-Paap F-statistic for weak instruments. ***$p < 0.01$, **$p < 0.05$ and *$p < 0.10$.

Table 8: Results for Agricultural Productivity in 1920

	Coffee Production Per Capita (1)	Rice Production Per Capita (2)	Maize Production Per Capita (3)
Panel A: OLS			
Norm. Respiratory Deaths	0.003	-0.014	0.007
	(0.008)	(0.009)	(0.025)
Panel B: 2SLS			
Norm. Respiratory Deaths	-0.067**	-0.137***	-0.295***
	(0.031)	(0.043)	(0.101)
Region FEs	Yes	Yes	Yes
Region-level controls	Yes	Yes	Yes
Initial period controls from 1872	Yes	Yes	Yes
K-P F-stat	22.61	24.42	23.59
p-value for the F-Stat	[0.001]	[0.001]	[0.001]
Mean of dependent variable	0.319	0.289	1.200
Observations	383	401	404

Notes: The table shows the OLS results (Panel A) and 2SLS results (in Panel B with IVResp as instrument). Standard errors reported in parentheses are clustered at the regional level. We use the variation for the period 1917-1919 to evaluate the effects on 1920 productivity measures. Region-level controls include the geographical characteristicts (altitude, latitude, longitude), distance to capital city, the literacy rates for male and female in 1920, a measure for frost intensity (a dummy that takes a value of one if more than three frost periods were reported). The initial controls of 1872 include the share of foreigners, share of people who were slaves, the share of employment in agriculture, manufacturing, and in services/retail, population density, the share of people of race "Branca". The K-P F-stat refers to the Kleibergen-Paap F-statistic for weak instruments. ***$p < 0.01$, **$p < 0.05$ and *$p < 0.10$.

Table 9: Literacy Rates by age groups in 1940

	20-29 years old		**30-39 years old**	
	Male (1)	Female (2)	Male (3)	Female (4)
Panel A: OLS				
Norm. Respiratory Deaths	0.010**	0.004*	0.002	-0.003
	(0.004)	(0.002)	(0.008)	(0.007)
Panel B: 2LS				
Norm. Respiratory Deaths	0.046**	-0.047***	0.054**	0.023
	(0.022)	(0.017)	(0.021)	(0.025)
Region-level Controls	Yes	Yes	Yes	Yes
Initial period controls from 1872	Yes	Yes	Yes	Yes
Geographical controls	Yes	Yes	Yes	Yes
K-P F-stat	12.45	11.20	8.90	6.83
p-value for the F-Stat	[0.002]	[0.003]	[0.006]	[0.014]
Mean of Dependent Variable	0.588	0.414	0.584	0.322
Observations	506	506	506	506

Notes: The table reports the OLS results in Panel A and the 2SLS results in Panel B. We use observations for respiratory death rates for the years 1917-1920 only. The instruments used are October's rain and October's precipitation. Standard errors reported in parentheses are clustered at the mesoregion level. The initial controls of 1872 include altitude, the normalized number of people with a mental and physical disease, the literacy rate, the share of foreigners, the share of workers in agriculture, manufacturing and services, the share of people who were slaves, the share of people of race "Branca" (white), population density. We also control for the share of total municipality expenditure allocated to education and

health in 1940; the number of teachers per person aged 10 and plus (teachers/Pop. Aged 10 above)*1000, the share of total education expenditure allocated to school supplies and also to the payment of teachers in 1940. All regressions include controls for the share of male and female foreigners in 1940. For the age groups 10-19 and 20-29, we also control for the literacy rate of adults aged 30 and above. The K-P F-stat refers to the Kleibergen-Paap F-statistic for weak instruments. ***$p < 0.01$, **$p < 0.05$ and *$p < 0.10$.

Table 10: Literacy Rates and Instruction, by Age Group and Gender in 1940

	Literacy Rate: Age 5 +		**Literacy Rate: Age 18 +**	
	Male	Female	Male	Female
	(1)	(2)	(3)	(4)
Panel A: OLS				
Norm. Respiratory Deaths	0.012**	0.015**	0.013**	0.001*
	(0.006)	(0.005)	(0.005)	(0.005)
Panel B: 2LS				
Norm. Respiratory Deaths	0.081***	0.030	0.067***	0.009
	(0.022)	(0.022)	(0.024)	(0.027)
Region-level Controls	Yes	Yes	Yes	Yes
Initial period controls from 1872	Yes	Yes	Yes	Yes
Geographical controls	Yes	Yes	Yes	Yes
K-P F-stat	7.77	7.77	7.77	7.77
p-value for the F-Stat	[0.008]	[0.008]	[0.008]	[0.008]
Mean of the dependent variable	0.489	0.348	0.521	0.322
Observations	674	674	674	674

Notes: The table reports the OLS results in Panel A and the 2SLS results in Panel B. The instruments used are October's rain and October's precipitation. Standard errors reported in parentheses are clustered at the mesoregion level. The initial controls of 1872 include altitude, the normalized number of people with a mental and physical disease, the literacy rate, the share of foreigners, the share of workers in agriculture, manufacturing and services, the share of people who were slaves, the share of people of race "Branca" (white), population density. We also control for the share of total municipality expenditure allocated to education and health in 1940; the number of teachers per person aged 10 and plus (teachers/Pop. Aged 10 above)*1000. We use observations for the years 1917-1920 only. The K-P F-stat refers to the Kleibergen-Paap F-statistic for weak instruments. ***$p < 0.01$, **$p < 0.05$ and *$p < 0.10$.

Table 11: In-Patient Hospital Admissions (per 1000) in 1940

	Hospital (1)	Asylum (2)
Panel A: OLS		
Norm. Respiratory Deaths	-0.032	-0.090
	(0.073)	(0.143)
Panel B: 2SLS		
Norm. Respiratory Deaths	1.291**	0.064
	(0.650)	(1.107)
Region FEs	Yes	Yes
Region-level controls	Yes	Yes
Geographical controls	Yes	Yes
K-P F-stat	10.37	0.90
p-value for the F-Stat	[0.000]	[0.000]
Mean of Dependent Variable	3.909	5.967
Observations	287	211

Notes: We use the variation for the period 1917-1920 to evaluate the effects on 1940 Inpatient Hospital Admissions. Standard errors reported in parentheses are clustered at the micro regional level. The controls include altitude, latitude, longitude, distance to the capital city, a dummy for the presence of railway. We also control for per capita total expenditure on health and the share of total health expenditure devoted to hospitals in 1940, and the number of hospital and nursing homes available. The K-P F-stat refers to the Kleibergen-Paap F-statistic for weak instruments. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$.

Table 12: Results for Aggregate Measures for the Primary Sector in 1940

	Value per Employee		Value per Establishment	
	(1)	(2)	(3)	(4)
Panel A: OLS				
Norm. Respiratory Deaths	0.013	-0.037	0.064*	-0.014
	(0.029)	(0.035)	(0.033)	(0.042)
Panel B: 2SLS				
Norm. Respiratory Deaths	-0.502*	-0.448***	-0.241	-0.472***
	(0.286)	(0.172)	(0.311)	(0.170)
Region FEs	Yes	Yes	Yes	Yes
Region-level controls	No	Yes	No	Yes
Initial period controls from 1872s	No	Yes	No	Yes
K-P F-stat	6.42	8.70	6.42	8.70
p-value for the F-Stat	[0.015]	[0.005]	[0.015]	[0.005]
Mean of dependent variable (in log)	6.781	6.781	8.839	8.839
Observations	736	468	736	468

Notes: The table shows the 2SLS results (with IVResp as instrument). Standard errors reported are clustered at the microregion level. We use the variation for the period 1917-1920 to evaluate the effects on 1940 agricultural productivity. In Columns (1) and (3), we consider the regressions the value of yield per employee and per establishment with only region FE. In Columns (2) and (4), we add the controls including altitude, latitude, longitude, distance to the capital city. We also control for the 1940 literacy rates for male and female of age 5 and above, and for the 1920 share of foreigners. Further, we include several other 1940 controls that could affect productivity. These controls include the expenditure per establishment by agricultural, farming and livestock establishments; the number of plows per establishment, the amount of expenditure on salaries, acquisition of machinery and of new animals, and on fertilizers per establishment. The initial controls of 1872 include population density, the shares of employment in agriculture, manufacturing, retail/services, the share of people of race "Branca", the share of people who were slaves. The K-P F-stat refers to the Kleibergen-Paap

F-statistic for weak instruments. ***$p < 0.01$, **$p < 0.05$ and *$p < 0.10$.

Table 13: Test of Sample Mean Differences in Literacy Rates for Districts in 1918

	Respiratory Death Rates					
	Below Median		*Above Median*			
	Mean	S.Dev	Mean	S.Dev	Diff.	p-value
Literacy Rates:						
Brazilian						
Male	0.257	0.059	0.249	0.007	0.007	0.480
Female	0.174	0.060	0.160	0.007	0.014	0.174
Total	0.217	0.059	0.206	0.073	0.011	0.290
Foreigners						
Male	0.562	0.153	0.524	0.149	0.038	0.083
Female	0.305	0.143	0.271	0.179	0.035	0.155
Total	0.461	0.152	0.425	0.155	0.035	0.120

Table 14: Falsification Test Using 1915 Deaths from Respiratory Infections

	1920	**1920**			**1940**			
	Sex Ratios	Literacy Rate: All Ages			Literacy Rate (5+)		Literacy Rate (18+)	
	(At Birth)	Male	Female	Total	Male	Female	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: OLS								
Norm. Respiratory Deaths	-0.028	0.008	0.009*	0.009*	0.079	0.011	0.010	0.011
	(0.050)	(0.005)	(0.005)	(0.004)	(0.008)	(0.008)	(0.008)	(0.006)
Panel B: 2SLS								
Norm. Respiratory Deaths	-0.179	0.038	0.041	0.038	0.036	0.020	0.0685	0.0324
	(0.440)	(0.043)	(0.045)	(0.043)	(0.035)	(0.028)	(0.044)	(0.027)
Region FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Initial period controls from 1872	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
K-P F-stat	3.61	2.15	2.15	2.15	2.52	2.52	2.52	2.52
Mean of the dependent variable	1.22	0.28	0.18	0.23	0.49	0.35	0.52	0.33
Observations	124	131	131	131	119	119	119	119

Notes: We use the same controls, fixed-effects and instruments as for the original regressions showing the impact of respiratory deaths during the pandemic years on the different outcomes of interest. ***$p < 0.01$, **$p < 0.05$ and *$p < 0.10$.

Table 15: Robustness Check for the Baseline Weight of the Constructed Instrument

	1915		**1920**	**1940**
	Still Births (Per 1000 Live Births)	Log Wages (Agricultural)	Productivity (coffee) (per establishment)	Value of primary sector (per establishment)
	(1)	(2)	(3)	(4)
Baseline Weight ($\bar{s}_{d1915}$)	22.014	0.728	1.743	-0.879
	(24.883)	(1.070)	(22.409)	(1.360)
Region FEs	Yes	Yes	Yes	Yes
Region-level controls	Yes	Yes	Yes	Yes
Initial period controls from 1872	Yes	Yes	Yes	Yes
Mean of the dependent variable	44.19	11.39	14.71	8.82
Observations	157	85	499	553

Notes: We use the same geographical controls and region-level initial controls as for the original regressions showing the impact of respiratory deaths during the pandemic years on the different outcomes of interest. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$.

Table 16: Impact of October's Temperature and Precipitation on 1920 and 1940 Outcomes

	1920				1940			
	Sex Ratios at birth		Log still births		Hospital Admissions		Literacy Rates	
							Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
October's Mean Temperature	-0.026	-0.041	0.013	0.016	0.030	0.025	0.004	0.002
	(0.025)	(0.042)	(0.011)	(0.011)	(0.059)	(0.070)	(0.004)	(0.005)
October's Mean Precipitation	0.109	0.002	-0.000	-0.000	0.020	0.019*	0.001	0.000
	(0.009)	(0.015)	(0.004)	(0.005)	(0.008)	(0.009)	(0.001)	(0.002)
Region FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Initial period controls from 1872	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	500	105	490	164	379	96	553	553
R-squared	0.466	0.571	0.253	0.255	0.530	0.532	0.460	0.447

Notes: The table shows the OLS regressions for the period 1917-1920 in columns (1), (3), (5) and (7), and for the year 1918 only in columns (2), (4), (6) and (8). Meso-region fixed effects are used in all regressions. The controls include altitude, latitude, longitude, railway dummy, distance to capital, and all the initial period controls from 1872 used previously in the analysis. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$.

A1 Appendix

In this section, we aim at estimating the effects of the 1918 Influenza Pandemic on specific cohorts using the IPUMS individual-level data for Sao Paulo for the years 1960, 1970, 1980. In Table A1 and Table A2, we show the cohort regressions for the following baseline specification used in Almond (2006):

$$y_i^c = \beta_0 + \beta_1 I(yob = 1919) + \beta_2 yob + \beta_3 yob^2 + \epsilon_i$$

where y_i^c denotes the census outcome for individual "i" in year 1960, 1970, and 1980. Our two main outcomes of interest pertaining to education are (i) whether the individual has completed high school or more, and (ii) whether the individual is literate. yob denotes the year of birth. We restrict the sample to those born between 1912 to 1922 and characterize the 1919 birth cohort as the one that experienced the greatest exposure to the pandemic (as in Almond (2006)). In other words, those born in 1919 are the "influenza treatment" group. We report the estimates for men, women, and non-whites, and focus on β_1-our estimate of interest, which as Almond explains, measures the departure of outcomes for the 1919 birth cohort from the quadratic cohort trend. While we are able to construct outcomes which closely mirror what Almond used, we are unable to have similar sets of controls and other outcomes of interest which could be used to lend further strength to the hypothesis that the 1919 birth cohort was negatively affected in several other ways. Our results in this setting suggest that there are no effects on high school completion for all groups considered except for men in 1960 (with the effect statistically significant at the 10% level only) and that there are possible selection effects on literacy rates (which mostly go away by 1980) as evidenced by the positive coefficients for men and women in 1960 and 1970 in Table A1.

Before proceeding with our region-level analysis of the effects of the Pandemic, we make a second attempt at using individual-level data by using Beach et al. (2018) regression analysis. They include individual and parental controls in addition to the cohort that Almond had, and consider the incidence of the flu at a more disaggregated level (cities).

DD model: $y_{ibc} = \alpha_0 + \beta_b + \gamma_c + \Sigma_{b=1915}^{1919} \delta_b 1[birthyear = b] \times flu_{1918,c} + \Gamma X' + \epsilon_i$

We provide a simpler version of their model and assign the same aggregate value of influenza-related death rates in 1918 to people in the census for Sao Paulo only. We focus once again only on education and literacy outcomes. We restrict the sample to those born between 1915 and 1919. The results are presented in Tables A3 and A4.

Table A1: Regression Results for Literacy Rates, Sao Paulo: 1960, 1970 and 1980

	Men			Women			Non-Whites	
	1960	1970	1980	1960	1970	1980	1960	1980
I(yob=1919)	0.025***	0.058***	0.015	0.033**	0.043***	0.015	-0.011	0.018
	(0.009)	(0.009)	(0.010)	(0.013)	(0.011)	(0.012)	(0.028)	(0.027)
yob	2.368**	-2.558**	-0.220	0.831	0.115	-5.944***	1.307	-6.103*
	(0.963)	(1.148)	(1.336)	(1.290)	(1.350)	(1.550)	(2.853)	(3.388)
yob2	-0.001**	0.001**	0.000	-0.000	-0.000	0.002***	-0.000	0.002*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Constant	-2,271.701**	2,450.348**	204.805	-802.201	-113.941	5,691.090***	-1,261.366	5,837.162*
	(923.417)	(1,100.310)	(1,280.369)	(1,236.735)	(1,293.861)	(1,485.584)	(2,735.046)	(3,248.281)
Observations	28,509	25,874	20,031	27,906	26,675	22,577	5,931	4,699
R-squared	0.001	0.002	0.003	0.002	0.002	0.003	0.004	0.009

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$.

Table A2: Regression Results for Literacy Rates, Sao Paulo: 1960, 1970 and 1980

	Dependent Variable: Literacy Rates					
	Men			Women		
	1960	1970	1980	1960	1970	1980
yob	0.018***	0.010***	0.005	0.024***	0.019***	0.014***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Respiratory Deaths*yob1917	-2.179	-1.842	0.412	-3.043*	-5.463***	-3.932**
	(1.495)	(1.549)	(1.901)	(1.754)	(1.787)	(1.965)
Respiratory Deaths*yob1918	-6.155***	-4.515***	-0.428	-4.977**	-4.260**	-4.968**
	(1.674)	(1.744)	(2.127)	(1.946)	(1.974)	(2.200)
Respiratory Deaths*yob1919	-5.840***	0.922	1.944	-5.016**	-3.541	-3.656
	(1.936)	(1.928)	(2.398)	(2.301)	(2.244)	(2.488)
Constant	-33.850***	-18.309***	-8.673	-46.142***	-36.583***	-26.138***
	(4.445)	(4.523)	(5.720)	(5.032)	(5.105)	(5.818)
Include controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	23,588	22,051	16,161	23,194	22,711	18,252
R-squared	0.061	0.085	0.091	0.128	0.147	0.137

Notes: Controls include indicator for race (white), whether anyone has completed high school in the household, whether individual is resident in the city, number of own family members in the household, number of own children in the household, number of own children under 5 years in the household, whether married and total income. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$.

Table A3: Differences in Sample Means for Selected Characteristics for Districts With Above and Below Influenza-Related Deaths during 1917-1920

	Below Median		Above Median			
Variables	Mean	Std Dev	Mean	Std Dev	Difference	
Municipality Expenditure (per capita)						
Cleaning and Waste Disposal	258.89	611.61	152.40	125.80	106.49	***
Maintenance of Sanitary Conditions	110.12	106.07	110.59	124.29	-0.47	
Geographic and Climate						
Altitude (Meters)	624.13	189.49	602.74	197.22	16.39	
Latitude	-22.23	1.05	-22.65	1.09	0.42	***
Longitude	-47.88	1.17	-47.63	1.39	-0.25	***
Distance to Capital	260.06	108.58	209.41	112.51	50.65	***
October's Temperature	20.61	1.98	19.91	2.10	0.70	***
October's Precipitation	124.85	7.54	123.67	9.18	1.18	**
Demographic						
Overall Median Age	20.75	2.92	20.62	2.96	0.13	
Population Density	9.56	6.53	10.62	8.38	-1.06	**
Dependency Ratio	88.91	21.23	89.70	21.66	-0.79	
Sex Ratio	109.93	6.77	108.66	8.32	1.27	**
Share of Slaves	0.15	0.07	0.15	0.08	0.00	
Literacy Rate	0.22	0.11	0.21	0.10	0.02	**
Economic						
Share of Employment in Agriculture	0.64	0.09	0.65	0.10	-0.00	
Share of Employment in Manufacturing	0.13	0.04	0.13	0.04	0.01	**
Share of Employment in Services	0.22	0.09	0.22	0.09	-0.11	
Health						
Physical or mental disease	11.07	5.19	10.80	5.50	0.28	
Doctors	0.22	0.28	0.18	0.24	0.03	*
Chemists	0.18	0.17	0.17	0.17	0.02	
Midwives	0.27	0.33	0.25	0.35	0.02	

Notes: This table shows the summary statistics and sample means differences for selected initial characteristics. The municipality expenditure per capita is for 1910, the geographic and climate variables are contemporary (1917-1920) and all other initial conditions are drawn from the 1872 census. The influenza-related deaths include deaths caused by respiratory and unknown causes. The median death rate is 8.11. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$.