The 1918 Influenza Pandemic and the Fetal Origins Hypothesis: Evidence from Linked Data

Preliminary Draft
June 16, 2017

Brian Beach
The College of William & Mary

Joseph Ferrie
Northwestern University

Martin Saavedra
Oberlin College

Abstract: Almond (2006) argues that in-utero exposure to the 1918 influenza pandemic lowered socioeconomic status in adulthood, whereas Brown and Duncan (2016) find that the effect disappears after controlling for parental characteristics of the 1919 birth cohort. We link microdata from the 1920 Census to WWII enlistment records and city-level influenza data. The result is a dataset with much more precisely measured influenza exposure and parental characteristics. We find results consistent with the Fetal Origins Hypothesis. In the absence of the pandemic, the 1919 birth cohort would have been more likely to graduate from high school and would have obtained more years of schooling. We find no evidence that in-utero exposure to influenza affected heights or weights.

JEL codes: I10; J13, N32

INTRODUCTION

To what extent does early-life exposure to disease and deprivation affect adult outcomes? Barker (1995) posits that malnutrition while in utero increases susceptibility to heart disease in adulthood. In a highly influential paper, Almond (2006) treats the 1918 Spanish Flu as a natural experiment to test this hypothesis. Specifically, he compares the adult outcomes of birth cohorts that were exposed to the flu in utero to adjacent birth cohorts that were not exposed. Almond finds that in utero exposure to the Spanish Flu decreased health in adulthood and that it also lowered educational attainment, income, and socioeconomic status (Almond, 2006; Almond and Mazumder, 2005). Economists have applied variations of this identification strategy to assess the implications of early-life exposure to disease on individual outcomes as well as overall economic development. The majority of these studies yield consistent results: early-life exposure to disease and deprivation lowers socioeconomic status and health in adulthood, which has important implications for development.[1]

The critical assumption in this literature is that in-utero exposure is the only determinant of human capital that systematically varies between birth cohorts. In a recent paper, however, Brown and Thomas (2016) cast doubt on this assumption. Brown and Thomas draw on data from the 1920 and 1930 US censuses to test whether parents whose children were in utero during the Spanish Flu were observably different from parents whose children were not exposed. The authors find that parents of the exposed cohort were less likely to be literate, were less likely to be white, and were of lower socioeconomic status. Brown and Thomas suggest that the reason for these differences is that the Spanish Flu coincided with World War I (WWI) and that WWI veterans were positively selected from the overall population. Importantly, Brown and Thomas show that many of Almond's findings are not robust to controlling for parental characteristics.

This study takes Brown and Thomas' criticisms seriously. However, one explanation for the lack of robustness to the inclusion of parental controls is that both Almond (2006) and Brown and Thomas (2016) rely on aggregate state-level data to perform their analysis. As we show, there was substantial variation in flu exposure within states and there was also variation in parental characteristics. Thus, any analysis that simply studies average flu exposure and controls for average parental characteristics will be subject to substantial measurement error.

[1] See Almond and Currie (2011) for a review of this literature. Some notable examples from roughly the same time period as the flu pandemic include Bleakley (2007) on hookworm eradication, Barreca (2010) on exposure to malaria, and Beach et al. (2016) on early-life exposure to typhoid fever.

In this paper, we dramatically reduce measurement error by linking individual administrative records over time. Specifically, we link young adult males from World War II enlistment records to their corresponding record in the 1920 census. This allows us to identify their city of residence in 1920, which we interpret as their in-utero environment. Because the individuals are observed with their parents in the census, we collect (and subsequently control for) parental characteristics such as mother's literacy, father's literacy, and whether the family owned or rented their home. We then digitize annual city-level disease data to construct a measure of Spanish Flu intensity. This allows us to adopt a differences-in-differences identification strategy that includes birth cohort fixed effects, as our identification comes from relative flu intensity. Our identification strategy allows us to estimate separate treatment effects for each birth cohort, thus, allowing us to disentangle the effects of prenatal exposure to influenza from exposure during early childhood.

Our results are twofold. First, as in Brown and Thomas (2016) we find some evidence that parents of the exposed cohort were observably different than parents of unexposed cohorts. However, even after including birth cohort fixed effects and controlling for parental characteristics, we continue to find evidence that is consistent with Almond (2006). Specifically, we see that increased exposure to the Spanish Flu lowered educational attainment, decreasing the likelihood that an individual would graduate high school or college. For a subset of our records, we observe the results from their Army General Classification Test (AGCT), which is a reasonable proxy for intelligence. While we find evidence that increased exposure lowered AGCT scores, the standard errors are relatively large and so we cannot reject that they are different from zero at standard levels of significance. Adult height and weight, however, appear to be largely unaffected by increased exposure to the Spanish Flu while in utero.

MEASURING THE LONG-RUN EFFECTS OF IN UTERO FLU EXPOSURE

Constructing a linked dataset

To assess the extent to which in utero flu exposure affected long-run outcomes, we construct a dataset that links individuals across administrative records. We begin with the WWII enlistment records, which were digitized by the National Archives and Records Administration.[2] These enlistment records span from 1938 to 1943 and cover over ten million men. It is important to

[2] Records are available at http://aad.archives.gov/aad/fielded-search.jsp?dt=893.

note that all men between the ages of 18 and 45 were required to register with the selective service, and so the sample is more representative than one might initially think.[3] The exception to this is that registrants could be rejected for failing to meet minimum education or physical standards. Failing to observe rejected applicants, however, would likely bias us against finding evidence that in-utero flu exposure impaired human capital development.

The enlistment records include the following information: name, race, year of birth, state of birth, years of secondary and post-secondary education, height, and weight.[4] In order to link these records to the 1920 census, we draw primarily on the digitized name, year of birth, and state of birth variables. We focus our attention on individuals that would have been between the ages 0 and 10 at the time of census enumeration (i.e. the 1909-1919 birth cohorts). The reason for this age restriction is that it increases the likelihood that the city of residence in the 1920 census is the same as the city of residence while in utero.

Our linking procedure closely follows the existing literature (e.g., Long and Ferrie, 2013; Aizer et al., 2016; Beach et al., 2016).[5] Specifically, we begin by standardizing all given names (e.g., "Ed" and "Eddie" would be recoded as "Edward"). Once names are standardized we, merge individuals from the enlistment records to the 1920 census based on the following information: year of birth (plus or minus two years), first initial of standardized given name, the Soundex (see discussion below) of the last name, middle initial (if present in both sets of records), race, and state of birth.[6] We allow year of birth to vary by up to two years to accommodate the fact that the information comes from two different sources (the year of birth reported in 1920 likely comes from a parent, while the year of birth reported in the enlistment records comes from the individual). Thus, to the extent that numeracy differed, there may be

[3] Of course, representativeness refers to a representative sample of males. While female records do exist, women were not required to register and so we might worry about whether the sample is representative. Because of this, we do not attempt to link female records. Failing to link women is in fact common practice in this literature (See for instance, Aizer et al., 2016 and Long and Ferrie, 2013). This is because women tend to change their name when they get married, and so without additional information on their maiden name, it is impossible to find their childhood administrative records.

[4] Between March and June of 1943, the individual's AGCT score, a reasonable proxy for intelligence, was entered in place of weight (Ferrie et al., 2012). We will eventually exploit this for our analysis.

[5] Alternative linking algorithms are available; however, recent work by Bailey et al. (2017) suggests that the method we employ is among the best in terms of reducing instances of false matches.

[6] Soundex is a phonetic algorithm for indexing names based on how they are pronounced in English. This allows us to match names despite minor differences in spelling. This is important because, during this time period, census enumerators went door-to-door and recorded the information that was *spoken* to them.

disagreement between the two sets of records. This procedure allows us to link 595,550 of the enlistees to a record in the 1920 census.

Measuring flu exposure

The 1918 influenza pandemic remains the deadliest influenza pandemic in history. The pandemic claimed over 50 million lives (Johnson and Mueller, 2002); however, from an empiricist's perspective this unfortunate event also presents a unique opportunity to assess the link between in-utero disease exposure and adult outcomes. Two features of the pandemic are particularly useful from an identification perspective. First, the pandemic was quick and unanticipated. While influenza is generally worse in the winter, this pandemic appeared in September and spread across the United States within a month (Sydenstriker, 1918). This is useful as it alleviates concerns about selection into pregnancy. Second, influenza is spread through the air, which combined with the severity of the flu, undercuts concerns that exposure would be related to socioeconomic status.

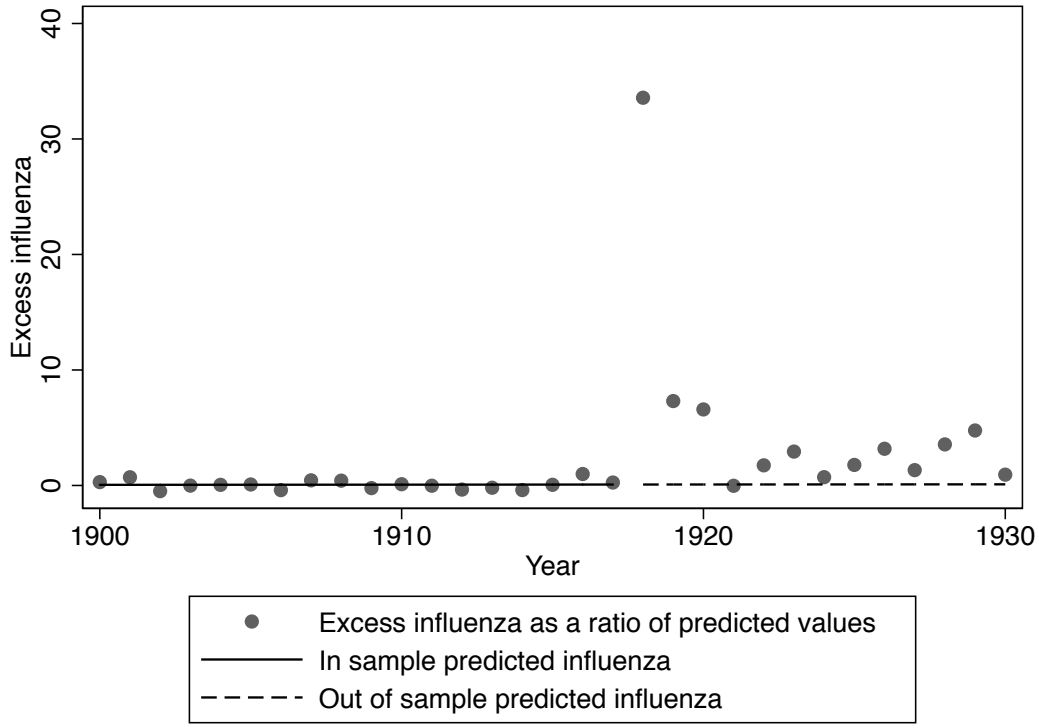
Our goal is to construct an independent variable that captures both of these features. Specifically, we are interested in a city-level measure of unanticipated flu intensity. To arrive at such a measure, we transcribe annual city-level flu mortality for the 1900 to 1930 period. We then regress the natural log of influenza fatality counts on a city-specific linear time trend for the 1900-1917 period. Taking the exponential of the predicted values from this regression yields a prediction of influenza fatalities in the absence of the pandemic for post-1917 years.[7] Our measure of excess influenza during the pandemic in birth year b in city c then becomes: $\text{flu}_{b,c} = \frac{\text{fatal}_{b,c} - \widehat{\text{fatal}}_{b,c}}{\text{fatal}_{b,c}}$, where $\widehat{\text{fatal}}_{b,c}$ is our predicted influenza fatalities and $\text{fatal}_{b,c}$ is the actual number of fatalities.

To implement this method, we transcribe annual city-level influenza mortality data spanning 1900-1930. In general, as time goes on more and more cities would be added to these mortality reports. There are 357 cities with at least 10 observations prior to 1918. For these cities, Figure 1 plots average excess influenza by year. The solid line is a line of best fit through

[7] The use of the natural logarithm ensures that the predicted number of influenza deaths is always greater than zero.

the years of 1900-1917 (the in sample period).[8] Excess influenza remains close to zero until 1918, during which approximately 35 influenza deaths occurred for every expected influenza death. Influenza deaths are higher during than the 1920s than would have been predicted using the pre-pandemic data, however, even 13 years after 1917, excess influenza deaths are not far above zero. It appears influenza reached its new steady state in 1921.

Figure 1: Mean excess influenza mortality



Notes: This figure reports the average excess influenza across all cities in our samples. Excess influenza as a ratio of predicted values is obtained by first regressing ln(influenza deaths) on a city-specific time trend. Taking the exponential of the residual yields the predicted mortality, and then excess influenza in city c during year t is calculated as $\frac{\text{flu}_{c,t}-\widehat{\text{flu}}_{c,t}}{\widehat{\text{flu}}_{c,t}}$. Influenza mortality data were transcribed from annual vital statistics reports.

One natural alternative measure of flu intensity is to use the influenza mortality rate (i.e., fatalities divided by population). However, because population is only observed in census years, the only way to obtain a population denominator is to interpolate between the 1910 and 1920

[8] Alternatively, we could have used data from the 1900-1917 and 1920-1930 years, omitting the years during which the pandemic occurred. However, if the pandemic had lingering effects on influenza rates or city population counts, then data from the 1920-1930 period would potentially be endogenous.

censuses, which seems problematic given the possible direct effect of influenza on population. Furthermore, influenza mortality rates would fail to capture the unanticipated dimension of the epidemic, as they do not take into account what the counterfactual influenza mortality rate would have been. For these reasons, we prefer our measure of flu intensity, however, we consider alternative measures as a robustness check and find similar results.

Table 1: Summary statistics

	Obs.	Mean	S.D.	Min	Max
1918 excess flu mortality	124,369	41.18	28.67	2.07	179.17
WWII records					
High school completion	124,369	0.55	0.49	0	1
College completion	124,369	0.09	0.29	0	1
Years of schooling	124,369	11.43	2.41	8	17
Height	96,369	68.26	2.76	60	78
Weight	95,321	153.2	23.86	105	300
AGCT score	3,541	109.09	19.04	42	159
Birth year	124,369	1915.35	2.97	1909	1919
1920 Census					
Mother can read	124,369	0.94	0.24	0	1
Father can read	124,369	0.96	0.2	0	1
Mother can write	124,369	0.94	0.25	0	1
Father can write	124,369	0.95	0.21	0	1
Owned home in 1920	124,369	0.31	0.46	0	1
Birth order	124,369	2.69	1.84	1	25

Next we merge excess influenza mortality in 1918 to our linked sample. 124,369 of our linked enlistees resided in a city for which we have influenza data and had both parents present at the time of enumeration. Table 1 produces summary statistics for our final dataset. The average individual in our sample experienced an excess influenza mortality ratio of 41, suggesting that for every anticipated influenza death there were an additional 41 unanticipated deaths. However, there is substantial variation in this variable, with some cities experiencing a ratio as small as 2 and others experiencing a ratio as high at 179. We also see that mean years of schooling was 11.4 with 55 percent of our sample completing high school and 9 percent completing college. In the appendix, we report summary statistics for the full linked sample.

There we see that relative to the full sample, the linked sample with flu exposure data is slightly more educated and their parents are less likely to have owned their home in 1920, perhaps reflecting that mortality data are only available for bigger cities, which may have a higher cost of living.

EMPIRICAL APPROACH

We estimate the following difference-in-differences model:

$$y_{ibc} = \alpha_0 + \beta_b + \gamma_c + \sum_{b=1915}^{1919} \delta_b 1[birthyear = b] \times \text{flu}_{1918,c} \quad + \Gamma X' + \varepsilon_i$$

where y_{ibc} is an adult outcome measured from the WWII records (education or AGCT score) for individual i, with birth year b, and living in city c in 1920. The model includes birth year fixed effects β_b and city fixed effects γ_c. The birth year fixed effects will account for the fact that the 1919 birth cohort may have systematically differed from neighboring birth cohorts, which addresses concerns raised in Brown and Duncan (2016). The city fixed effect will account for the fact that cities disproportionately struck by the pandemic may have systematically differed from cities that were relatively spared. For example, cities with greater pandemic severity may have been better connected to transportation networks. The variable $\text{flu}_{1918,c}$ is our measure of excess influenza as described in the previous section. The coefficients of interest are δ_b, which measures a separate treatment effect for each birth cohort. We estimate δ_b for birth cohorts from 1915-1919. The cohorts born before the 1919 act a placebo test, and if the 1918 pandemic primarily affected labor market outcomes of those in utero during the pandemic, we would expect $\delta_{1914} = \delta_{1915} = \cdots = \delta_{1918} = 0$.

The vector X' is a set of control variables that include race and the following 1920 characteristics: mother's literacy, father's literacy, mover/stay status, home ownership, and birth order. Standard errors are clustered at the 1920-city level. For binary dependent variables, we use a probit model in place of OLS and report average marginal effects.

In addition to the difference-in-differences model, we estimate two models that make use of excess influenza rates during years outside of the 1918 pandemic. First, we estimate the following OLS regression:

$$y_{ibc} = \alpha_0 + \beta_b + \gamma_c + \delta flu_{b-1,c} + \Gamma X' + \epsilon_{ibc}$$

where all variables are defined as before, except $flu_{b-1,c}$ is the excess influenza death rate in the individual's city of birth during the year before their birth (presumably when they were in utero). This model has the advantage of using variation in influenza from before the pandemic. This model is also more efficient, since there are fewer parameters and we gained influenza variation from years other than 1918. However, this model assumes that influenza exposure during stages other than the in utero period are unaffected by influenza exposure.

For the third specification, we run is a variation of the second, but do not make a parametric assumption regarding the functional form between in-utero influenza exposure and adult outcomes. Instead, we estimate the following semi-parametric equation using partial linear regression:

$$y_{ibc} = \alpha_0 + \beta_b + \gamma_c + f(flu_{b-1,c}) + \Gamma X' + \epsilon_{ibc}$$

where the function *f* is some continuous function that will be estimated using a local polynomial smooth. Allowing for some non-linearity in the influenza-outcomes relationship will allow for the fact that cities exposed to especially high levels of influenza may suffer from selective mortality.

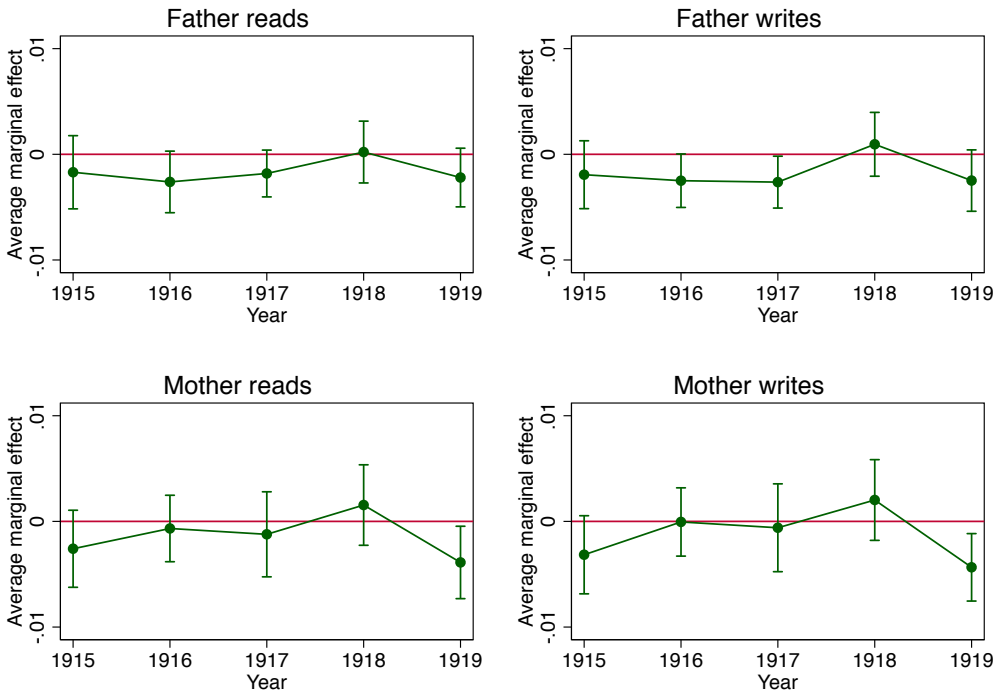
RESULTS

Parental characteristics

Figure 2 assesses whether cohorts exposed to the pandemic in utero had systematically different parents relative to unexposed cohorts. Specifically, we regress parental characteristics on birth year fixed effects, city fixed effects, and excess flu mortality interacted with dummies for the 1915-1919 birth cohorts. For ease of interpretation, we standardize our excess mortality variable before interacting it with our birth cohort fixed effects. We then plot the marginal effects from these probit regressions in Figure 2. In general, we see negative point estimates, with p-values

just over 0.05 suggesting that parents whose children were exposed to increased excess flu mortality in utero may have been less literate. This is consistent with Brown and Duncan (2016), and suggests that parental characteristics will be an important set of controls for analysis.

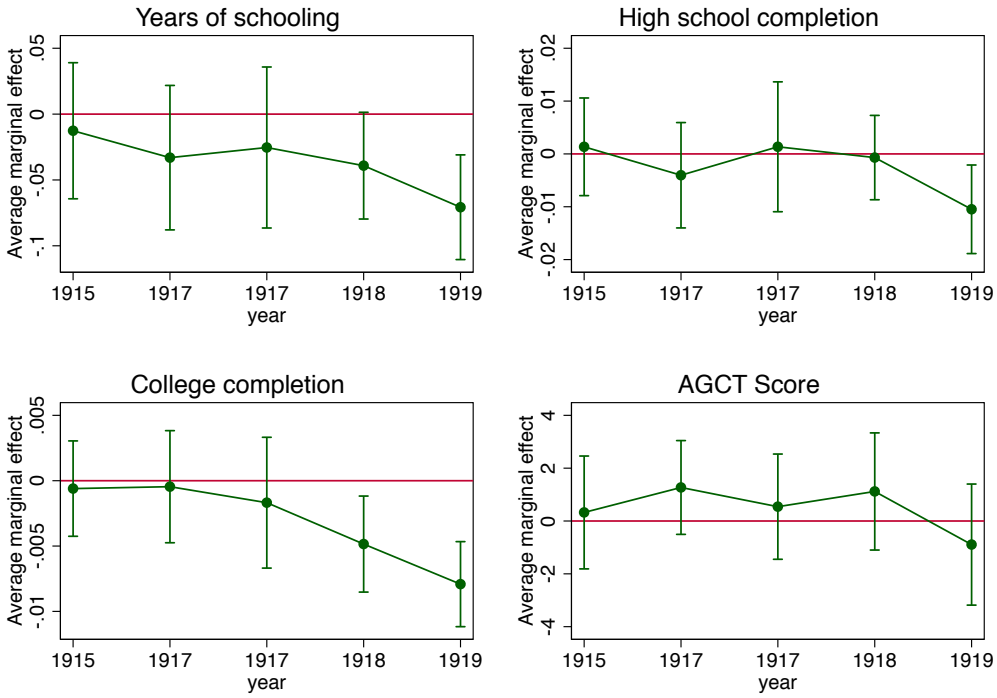
Figure 2: Assessing the validity of our empirical approach



Notes: Standard errors are clustered at the birth-city level. Full regression also includes birth cohort fixed effects and city fixed effects.

Adult outcomes

Figure 3: Excess flu exposure and educational outcomes



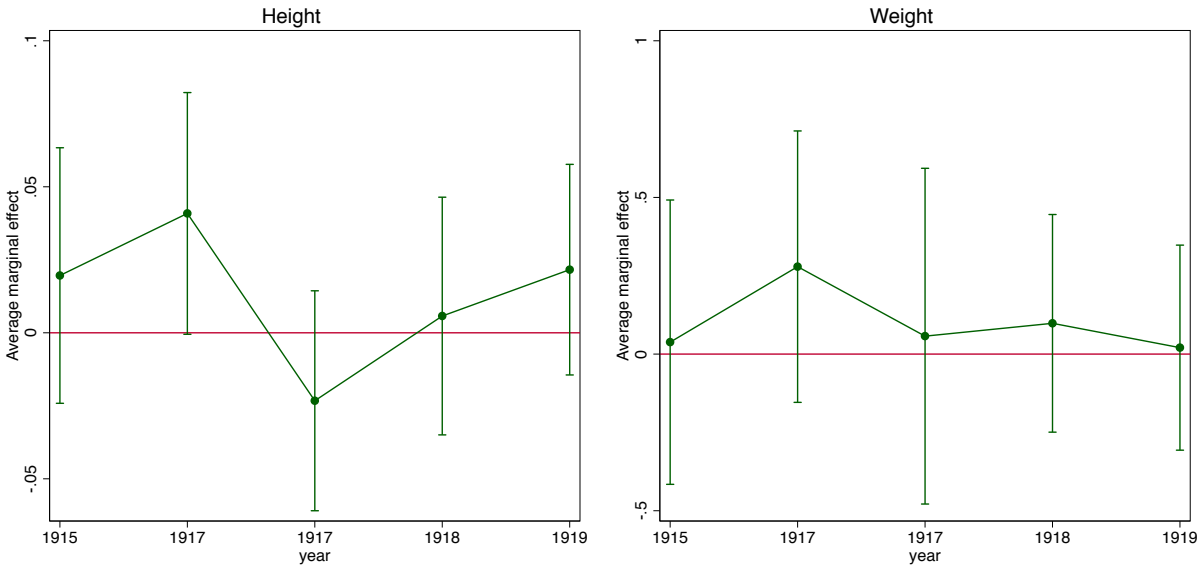
Notes: Standard errors are clustered at the birth-city level. Full regression also includes birth cohort fixed effects, city fixed effects. Regressions also control for parental characteristics (parental literacy and whether parents owned their home in 1920) as well as controlling for the individual's birth order and race.

The main education results are presented in Figure 3. As in Figure 2, we present point estimates and 95% confidence intervals for the cohort-specific treatment effects. Each of these regressions also controls for parental characteristics (parental literacy and whether parents owned their home in 1920) as well as controlling for the individual's birth order and race. The results indicate that increased in utero flu exposure decreased overall educational attainment. Specifically, increasing in utero flu exposure by one standard deviation would have reduced schooling by 0.075 years. For the same increase in exposure, the likelihood of completing high school and the likelihood of completing college both decrease by about 1 percentage point. Finally we show results on the AGCT score. Between March and June of 1943, the individual's AGCT score was entered in place of weight (Ferrie et al., 2012). Previous studies (e.g., Ferrie et

al., 2012; Aizer et al., 2016) have used this as a proxy for intelligence, we attempt to do the same; however, only 2,121 of our enlistees have this data available. Accordingly, we find a meaningful effect on AGCT score – increasing exposure by one standard deviation lowers AGCT performance by the score by 1 points (roughly 1 percent relative to the mean). However, the standard errors are quite large, reflecting the dramatically reduction in our sample size.

Next we consider whether in utero exposure affected height or weight, the two biological outcomes that are available in the enlistment records. These results are presented in Figure 3. We find no evidence that in utero exposure to the 1918 influenza pandemic affected weights or heights. It is worth noting that this could be because the pandemic did not affect heights or weights, or it could also be explained by the fact that rejected applicants (those that do not meet a minimum educational or physical standard) do not appear in our data. If the pandemic did decrease physical ability, but physical standards truncate our distribution, then this would bias us toward finding no effect.

Figure 4: Excess flu mortality and biological outcomes



Notes: Standard errors are clustered at the birth-city level. Full regression also includes birth cohort fixed effects, city fixed effects. Regressions also control for parental characteristics (parental literacy and whether parents owned their home in 1920) as well as controlling for the individual's birth order and race.

Ordinary Least Squares Estimates

In this section, we regress adult outcomes on excess influenza during an individual's birth city during the year before their birth, birth year fixed effects, city fixed effects, and control variables. This approach makes use of excess influenza rates for years before the 1918 pandemic, but does not include placebo tests for exposure at a later age. The results, presented in Table 2, suggest that excess influenza during the in utero period reduced years of schooling, high school graduation rates, and college graduation rates. These results are statistically significant at the 1% level regardless of whether we control for parental characteristics are not. The estimates are about 5% smaller when we control for parental characteristics. We find no evidence of an effect for height or weight. AGCT scores are estimated with much less precision since the sample size is smaller, but the estimated sign is negative. The results for AGCT scores are only statistically significant at the 10% level if we do not control for parental characteristics.

Table 2: Excess in-utero flu exposure and adult outcomes

	Panel a: With parental controls					
	(1) Years of Schooling	(2) High school	(3) College	(4) Height	(5) Weight	(6) AGCT
Excess Flu	-0.0022***	-0.0004***	-0.0003***	0.0007	-0.0006	-0.057
	(0.0005)	(0.0001)	(0.0001)	(0.0006)	(0.0052)	(0.0396)
Observations	122740	122740	122740	95071	94090	3490
	Panel b: Without parental controls					
Excess Flu	-0.0024***	-0.0054***	-0.0003***	0.0006	-0.0005	-0.0703*
	(0.0005)	(0.0001)	(0.0001)	(0.0006)	(0.0053)	(0.0381)
Observations	122740	122740	122740	95071	94090	3490

Standard errors (clustered at the birth-city level) are reported in parentheses.
* p<0.10; ** p<0.05; *** p<0.01.

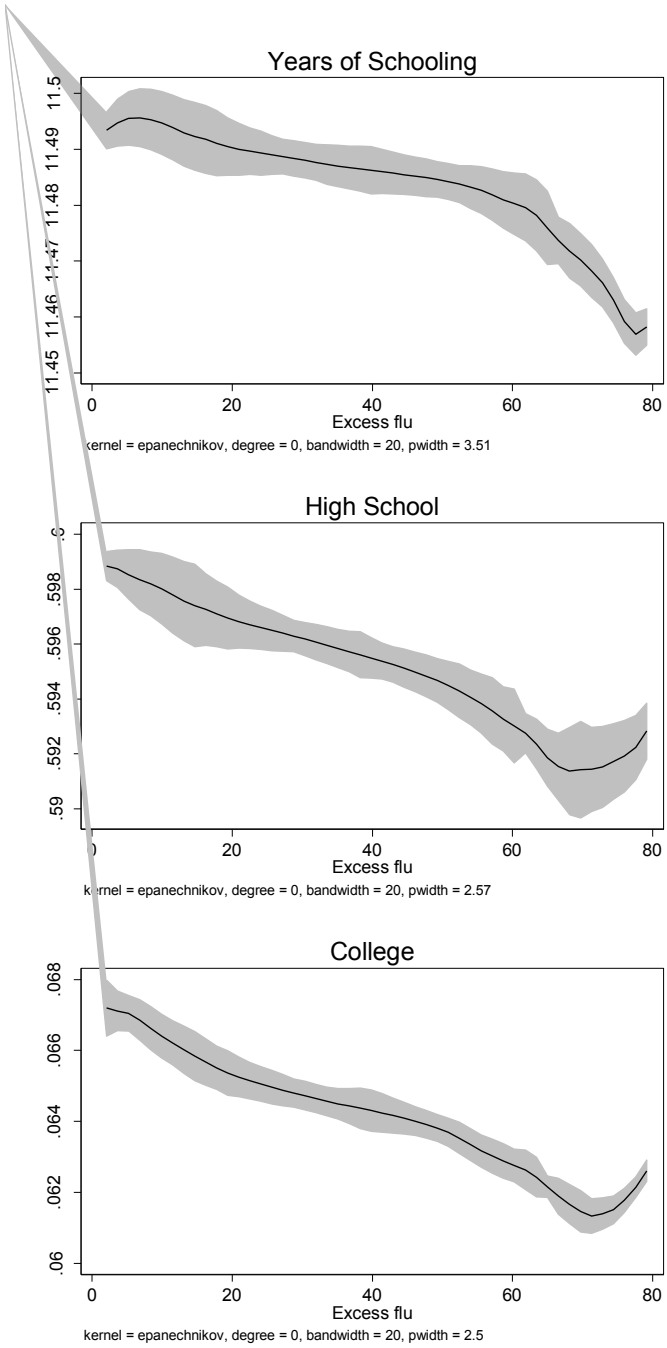
Notes: All regressions include birth-cohort fixed effects, birth-city fixed effects, birth order fixed effects and race fixed effects. Regressions with parental controls include indicators for parental literacy (whether mother can read, mother can write, father can read, and father can write) and homeownership status.

Semi-parametric Estimates

In this section, we present the semi-parametric results. For this analysis, we use partial linear regression by aggregating the data to the city-birth-cohort level. We first estimate the parametric part of the equation by sorting the data by excess influenza, and taking first differences. We then estimate the non-parametric part of the equation using a local polynomial smooth. This methodology is described in detail in Yatchew (2003).

Estimates of the *f* function are in Figure 5, and 95% confidence intervals are the shaded area. The semi-parametric estimates suggests that individuals exposed to higher levels of influenza obtained fewer years of schooling, were less likely to graduate from high school, and less likely to graduate from college. The relationship starts to reverse beyond 75 excess influenza deaths, implying that cities that were exposed to extremely high levels of influenza may have experienced selective mortality. The results suggest that males in utero in the worst cities affected by the pandemic would have had an additional 0.05 years of schooling, and would have been 1 percentage point more likely to graduate from high school and 0.6 percentage points more likely to graduate from college in the absence of the pandemic. Overall, it appears that a linear line would approximate the non-parametric line reasonably well.

Figure 5: Semi-parametric estimates of in-utero flu exposure



Notes: Each regression is estimated using partial linear regression, and parametrically controls for parental literacy, parental home ownership, birth order, 1920 city of residency, birth year, and race. The shaded region corresponds to 95% confidence intervals.

Selection on Unobservables

The previous subsections provide evidence in favor of the fetal origins hypothesis even after controlling for parental literacy, home ownership status, mover/stayer status, and birth order. Including controls from the 1920 census reduces estimates of the pandemic's effect, but only by small magnitudes. Many parental characteristics are still unobservable, and the inclusion of such controls could further reduce the effect of the pandemic if such controls were observed. In this subsection, we attempt to learn about the degree of selection of unobservable variables by analyzing selection on observable variables using the method of Oster (2016). This method provides us with two estimates. First, relative to selection on observable controls, how much selection on unobservables would there have to be for the estimate of the pandemic to be reduced to zero? Second, assuming that the degree of selection on observables is the same as the degree of selection on unobservables, what is the bias-adjusted estimate of the average treatment effect? Since observable covariates such as race, parental literacy, and home ownership likely have strongly correlated with adult educational attainment, we view this second estimate a lower bound on the treatment effect.

To use the Oster (2016) methodology, we must estimate two regressions: the full regression (the controlled model) and the regression without 1920 controls (the uncontrolled model). We include birth city and birth year fixed effects in both regressions, but the uncontrolled model excludes parental literacy dummies, birth order, mover/stay status, home ownership status, and race dummies. Changes in the betas are informative for how stable the estimates are, but these coefficient changes must be scaled by change in the R-squared. Additionally, we must specify a maximum R-squared that would be the R-squared if there were no unobservable variables excluded from the model. There are two reasons that the maximum R-squared could be less than one. First, the R-squared will be less than one if there is measurement error in the outcome variables. In our case, this would most likely occur if the two linked records are not true links. Second, there may be idiosyncratic components of educational attainment that are determined well after the 1920 Census. Consequently, even if we observed all determinants of educational attainment in the 1920 Census, we would get an R-squared less than one.

To determine the hypothetical maximum R-squared we would get if we observed all 1920 Census characteristics, we turn to twin studies. Ashenfelter and Rouse (1998) estimate that from a sample of twins that twin fixed effects can explain 60% of the variation in schooling. Since twins share the same birth cohort, birth city, parental characteristics (both observed and unobserved), and exposure to the same diseases in utero, we treat 60% as the maximum R-squared possible.

Taking our years of schooling estimates from Table 4, we get an estimate of -0.00218 with an R-squared of 0.115 for the controlled model. Running the sample regression without the 1920 controls or the race group controls yields as slightly larger estimate of -0.00231 with an R-squared of 0.061. The Oster adjustment suggests that selection on unobservables would have to 1.95 times larger than selection of observed variables for the estimated treatment effect to be zero. Additionally, the bias-adjusted estimate of beta reduces our estimate to -0.00106, suggesting that the pandemic reduced years of schooling for the 1919 birth cohort by about 0.5 months (mean excess influenza exposure for the 1919 birth cohort is approximately 42). We view this as the smallest estimate that is reasonable.

Discussion and Conclusion

A growing literature in economics uses epidemics as a natural experiment to assess whether in-utero exposure to disease and deprivation impairs human capital development. The seminal paper in this literature is Almond (2006), who studies the effects of the 1918 influenza pandemic. In a recent paper, however, Brown and Duncan (2016) argue that this experiment does not offer clean identification because parents of the 1919 birth cohort were systematically different. The reason for this is that the pandemic struck during WWI and WWI veterans were positively selected from the overall population. Furthermore, Brown and Duncan show that Almond's original results are not robust to controlling for parental characteristics.

In this paper, we note that measurement error could explain this lack of robustness. Specifically, Almond (2006) and Brown and Duncan (2016) both rely on aggregate state-level data. Flu intensity, however, varied both across and within states. To address this issue, we construct a new dataset linking WWII enlistees to 1920 census records. This allows us to identify

the city of residence (and thus in utero environment). Pairing these records with annual city-level mortality data allows us to construct a measure of flu intensity. This allows us to adopt a differences-in-differences strategy that includes birth cohort fixed effects, and thus relies on flu intensity as a source of variation. Results indicate that flu intensity (conditional on birth year) did not have an added effect on parental characteristics. In-utero exposure, however, did decrease educational attainment even after controlling for parental characteristics.

Moving forward we are in the process of linking 1930 census records to the WWII enlistment records. The benefit of using 1930 data is that it will allow for a richer set of parental control variables. Specifically, 1930 asks about father's veteran status and so we will be able to test the extent to which in-utero flu exposure was related to veteran status. Further, we will also be able to control for veteran status, which is the primary variable of interest in Brown and Thomas (2016). A second direction that we are exploring is constructing a linked dataset of brothers, as in Parman (3015). Parman (2015) constructs links brothers from the WWII enlistment records to census records in order to assess whether households with a child that was in utero during the pandemic respond by changing how they allocate resources among their children. By linking the full set of WWII records we would be able to also exploit our city-level data on flu intensity while also including household fixed effects. The inclusion of household fixed effects would go a long way in addressing concerns about selection bias, which would in turn offer much more precise estimates of the impact of in-utero flu exposure.

References

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney. 2016. "The long-run impact of cash transfers to poor families." *American Economic Review,* 106(4): 935-971.

Almond, Douglas. 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, Douglas, Janet Currie, and Valentina Duque. 2017. "Childhood Circumstances and Adult Outcomes: Act II." *National Bureau of Economic Research* working paper No. w23017,.

Almond, Douglas, and Bhashkar Mazumder. 2005. "The 1918 influenza pandemic and subsequent health outcomes: an analysis of SIPP data." *American Economic Review, 95* (2): 258-262.

Almond, Douglas, and Janet Currie. 2011. "Killing me softly: The fetal origins hypothesis." *The Journal of Economic Perspectives* 25(3): 153-172.

Ashenfelter, Orley, and Cecilia Rouse. 1998. "Income, schooling, and ability: Evidence from a new sample of identical twins." *The Quarterly Journal of Economics,* 113(1): 253-284.

Bailey, Martha, Morgan Henderson, and Catherine Massey. 2017. "How do automated linking methods perform? Evidence from the LIFE-M project." *University of Michigan, [Unpublished manuscript]*.

Barreca, Alan I. 2010. "The long-term economic impact of in utero and postnatal exposure to malaria." *Journal of Human Resources* 45(4): 865-892.

Beach, Brian, Joseph Ferrie, Martin Saavedra, and Werner Troesken. 2016. "Typhoid fever, water quality, and human capital formation." *The Journal of Economic History* 76(1): 41-75.

Brown, Ryan, and Duncan Thomas. 2016. "On the long term effects of the 1918 US influenza pandemic." *University of Colorado Denver [unpublished manuscript]* .

Case, Anne, and Christina Paxson. 2009. "Early life health and cognitive function in old age." *The American Economic Review* 99(2): 104.

Ferrie, Joseph, Karen Rolf, and Werner Troesken. 2012. "Cognitive disparities, lead plumbing, and water chemistry: Prior exposure to water-borne lead and intelligence test scores among World War Two US Army enlistees." *Economics & Human Biology,* 10(1): 98-111.

Johnson, Niall and Juergen Mueller. 2002. "Updating the accounts: global mortality of the 1918-1920" Spanish" influenza pandemic." *Bulletin of the History of Medicine,* 76(1): 105-115.

Long, Jason and Joseph Ferrie. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States since 1850." *American Economic Review*, 103(4): 1109-37.

Oster, Emily. 2016. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business and Economic Statistics*, forthcoming.

Parman, John. 2015. "Childhood health and sibling outcomes: Nurture reinforcing nature during the 1918 influenza pandemic." *Explorations in Economic History, 58: 22-43.*

Saavedra, Martin Hugo. "Early-Life Disease Exposure and Occupational Status: The Impact of Yellow Fever during the 19th Century." *Explorations in Economic History* (forthcoming).

Sydenstricker, Edgar. 1918. "Preliminary statistics of the influenza epidemic." *Public Health Reports (1896-1970)*: 2305-2321.

Yatchew, Adonis. 2003. "Semiparametric regression for the applied econometrician." *Cambridge University Press,*.

Appendix Table 1: Summary statistics for full linked sample

	Obs.	Mean	S.D.	Min	Max
1918 excess flu mortality	257,441	34.9	24.68	2.07	179.17
WWII records					
High school completion	829,922	0.48	0.49	0	1
College completion	829,922	0.07	0.26	0	1
Years of schooling	829,922	10.91	2.44	8	17
Height	621,084	68.46	2.73	60	78
Weight	613,309	152.71	23.32	105	300
AGCT score	28,920	103.67	21.36	40	160
Birth year	829,922	1915.27	2.99	1909	1919
1920 Census					
Mother can read	467,373	0.95	0.22	0	1
Father can read	469,419	0.95	0.22	0	1
Mother can write	467,146	0.94	0.23	0	1
Father can write	469,254	0.94	0.23	0	1
Owned home in 1920	465,117	0.42	0.49	0	1
Birth order	829,922	2.99	2	1	74