README for "The Intergenerational Effects of Early-Life Pollution Exposure" Replication

The code in this replication package constructs all tables and figures from a set of pre-merged analysis files combining air pollution data, confidential administrative records and internal use survey data. The process for requesting access to these confidential datasets and constructing the relevant analysis files is described below. Once a replicator has constructed the analysis file, the code in this replication package can be run in sequential order to create the exhibits.

Details on Analysis Data Construction

1) All the results in the paper use analysis datasets constructed from confidential administrative microdata linked with survey and Census data held at the U.S. Census Bureau. To gain access to the confidential microdata, you must be authorized to use this data under select internal US Census Bureau projects that are approved by Census and other partner agencies (either as US Census Bureau staff or approved SSS co-authors). Currently the data used to construct these analysis files is available under internal project DMS 7505723 and several closely related projects.

2) The project uses the following raw confidential data inputs

    a) Census Numident, vintages 2012 and 2017

    b) Decennial Census microdata + PIK crosswalks, 2000 and 2010

    c) American Community Survey microdata + PIK crosswalks, 2001-2021

    d) Master Address File Auxiliary Reference File, 2000-2021

3) In addition to confidential microdata, the project uses publicly available data sources:

    a) Monitor Level TSP exposure from EPA's AirNow API

    b) Temperature and precipitation data from ERA5 reanalysis

    c) CPI price deflator data from BLS

    d) County level Personal Income and Population from the BEA

4) To construct the "first generation analysis data" – referenced as "first_gen_analysis_data_jepmicro_acs.rda"—and the second generation analysis dataset– referenced as "second_gen_college_analysis_data_jpemicro.rda":

    a) Construct County-level annual TSP and in-utero (9 month average) exposure from the monitor data, weighting by the number of valid hourly observations per monitor

    b) Aggregate weather controls to the county-year or county-in utero period level

    c) Construct definitions of effective nonattainment of TSP NAAQS based on 1971 TSP values from 4a

    d) Identify all native-born individuals (PIKs) born 1960-1980 in the Census Numident

e) Identify all linkable children of the individuals in 4c in the 2000 and 2010 Decennial Census HDF files, and the 2001-2015 ACS one year files. Deduplicate by parent pik-child pik. This is the intergenerational linkage dataset.

f) Link all first generation individuals in the Numident to the Numident place of birth to county crosswalk, and then link TSP, weather controls and local economic controls by county and year of birth (and by in utero period). This is the first generation exposure dataset.

g) Link the first generation exposure dataset in 4f with the pooled 2005-2015 American Community Survey, retaining outcome variables (wages, employment status recode, disability) and relevant controls (age, race, sex, marital status). If a pik appears more than once, retain the observation with the latest survey year (if a pik appears more than once in a survey year, retain the response with the latest effective date using the rdate variable). This is now **the first generation analysis dataset**.

h) Link the first generation exposure dataset to the deduplicated parent-child links from the intergenerational linkage dataset. Retain all parent-child links, so the unit of observation is now parent-by-child.

i) Link the dataset from 4h with the pooled 2005-2021 ACS, retaining outcome variables (educational attainment, school attendance, disability status, wages, employment status recode), and relevant controls. If a pik appears more than once, retain the observation with the latest survey year (if a pik appears more than once in a survey year, retain the response with the latest effective date using the rdate variable). This is now the **second generation analysis dataset.**

Details on Tables and Figures Construction

1) Once the analysis datasets are constructed, the remainder of the code in this repository can then be run sequentially to create the exhibits in the paper. We list each script below.
2) Utility programs
   a) Child_outcomes_functions.R
   b) Rounding.r
3) Code to produce statistics for disclosure for main tables and figures
   a) Figure_1.R
   b) Figure_2.R
   c) Table_1.R
   d) Table_2.R
   e) Table_3.R
   f) Table_4.R
   g) Table_5.R
   h) Table_6.R

4) Code to produce statistics for disclosure for Appendix Tables and Figures
   a) Table_A1.R
   b) Table_A2.R
   c) Table_A3.R

d) Table_A4.R
e) Table_A5.R
f) Table_B1.R
g) Table_B2.R
h) Table_B3.R
i) Table_B4.R
j) Table_B5.R
k) Table_B6.R
l) Table_B7.R
m) Table_B8.R
n) Figure_A1.R
o) Figure_A2.R
p) Figure_B1.R
q) Figure_B2.R
r) Figure_B3.R
s) Figure_B4.R
t) Figure_B5.R

5) Code to produce figures from disclosed output
   a) Fig1.do
   b) Fig2.do
   c) FigA1.do
   d) FigA2.do
   e) FigB1.do
   f) FigB2.do
   g) FigB3.do
   h) FigB4.do
   i) FigB5.do