README for "Maternal and Infant Health Inequality: New Evidence from Linked Administrative Data"*

Kate Kennedy-Moulton† Sarah Miller‡ Petra Persson§ Maya Rossin-Slater¶
Laura R. Wherry|| Gloria Aldana**

June 6, 2025

1 Overview

This repository contains all the code needed to produce the main datasets and analysis results in the paper. No data can be made publicly available due to use of restricted Census FSRDC data as well as permission restrictions on the detailed and sensitive individual-level Swedish administrative data.

There are two main parts of the following programs that build the samples and create all gradients: those run on the Census FRSDC system and those run on the server where the Swedish administrative data is stored. These programs are produce the output that is then disclosed and subsequently formed into the tables and figures in the paper by Part 3 of the programs.

The included code was run using Stata (Version 18) and SAS (version 9.4).

The CA data building portion of the project takes three to four days to run depending on computational resource availability. The Swedish data building portion takes about a day to run. The analysis and graphing portions using these cleaned data takes less than an hour to run.

2 Data availability and sources

2.1 Details for obtaining access to each data source

The data used in this manuscript were obtained via several data use agreements described below.

*This research was conducted as a part of the U.S. Census Bureau's Evidence Building Project Series. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project P-7523134: CBDRB-FY22-CES018-005, CBDRB-FY22-CES018-012, CBDRB-FY22-CES018-016, CBDRB-FY22-420, CBDRB-FY23-0405, CBDRB-FY23-0464, CBDRB-FY25-0008, and CBDRB-FY25-0028).

†Columbia University. Email: kam2373@columbia.edu
‡University of Michigan Ross School of Business and NBER. Email: mille@umich.edu
§Stanford University, Research Institute of Industrial Economics, and NBER. Email: perssonp@stanford.edu
¶Stanford University School of Medicine, NBER, IZA. Email: mrossin@stanford.edu
||New York University Robert F. Wagner Graduate School of Public Service and NBER. Email: laura.wherry@nyu.edu
**US Census Bureau. Email: gloria.g.aldana@census.gov

Birth certificate date: Researchers interested in using the California birth certificate records held at the Census Bureau will require permission from the California Department of Public Health (CDPH) via an application process and an amendment to the existing agreement with the U.S. Census Bureau. Instructions for requesting access to vital statistics from CDPH are available here.

Linkage to Census-held data: After obtaining approval from the California Department of Public Health, researchers must request access to Census-held birth certificate data linked to each file from the U.S. Census Bureau to be accessed in the Federal Statistical Research Data Center (FSRDC). First, contact your FSRDC administrator to begin the process. Second, complete the application and/or follow the FSRDC administrator's guidance on how to apply. Since not all datasets may be listed online, you may need the FSRDC administrator to request them on your behalf.

Linkage to Internal Revenue Service (IRS) data: The US Census Bureau and the IRS must establish a data use agreement allowing the use of IRS data for the purpose of this replication.

Linkage to the Health Care Access and Information data: Researchers must obtain approval from the California Department of Health Care Access and Information and the California Department of Public Health to use the Linked Birth Files, which have information on hospitalizations, emergency department visits, and infant mortality linked to selected fields from the birth certificate records, as well as permission to link these files to the complete version of the California birth certificate records provided by the California Department of Public Health and to store these records in the Census FSRDC environment. Instructions for requesting access to these data files is available here.

Small Area Income and Povery Estimates (SAIPE): We obtained publicly available data from the SAIPE program for years 2005-2016. Specifically, we use data on CA county-level poverty rate and median income level. Copies of these raw data are provided in this package (data/raw/est*all.xls).

Statistics Sweden: Researchers interested in obtaining the data must first apply to the Ethical Review Authority ("Etikprövningsmyndigheten"). Following approval from the ERA, data from Statistics Sweden may be ordered.

National Board of Health and Welfare: Researchers interested in obtaining the data must first apply to the Ethical Review Authority ("Etikprövningsmyndigheten"). Following approval from the ERA, data from the National Board of Health and Welfare may be ordered.

Statistics Sweden assigns each individual a unique identification number that is a scrambled version (i.e., distinct from) the actual personal identification number used by the Swedish government. This unique identifier makes it possible to link data the sources described above. Specifically, Statistics Sweden sends the identifier to the National Board of Health and Welfare in order to obtain information from the health registers. Statistics Sweden and the National Board of Health and Welfare carry out all these steps. Thus, the data sets that are delivered to the researchers do not contain personal information such as the (true) personal identification number, or individuals' names and addresses. The data are delivered either to the servers maintained by Statistics Sweden through the MONA online servers (in which case the MONA data user agreement must be signed) or delivered directly to the researcher (in which case the data files must remain on a server within the European Union). All handling of the data must comply with all applicable Swedish laws.

Swedish CPI: The Official Statistics of Sweden publishes publicly available data on CPI with base years 1980 and 2010. We use data for years 2005-2016. A copy of this data is provided as part of this package (data/SE_data/CPI_Sweden_Oct2020.dta).

Table 1: Dataset List

Data Name	Years	Availability	Citation
California Birth records			
Annual Birth Certificate Data	2007-2016	By agreement only	California Department of Public Health
Maternal and infant personal identification key files	2007-2016	By agreement only	US Census Bureau
Paternal personal identification key files	2007-2016	By agreement only	US Census Bureau
Restricted-Use Census held data			
Decennial Census	2010	By agreement only	US Census Bureau
American Community Survey	2007-2021	By agreement only	US Census Bureau
Household composition key	2016-2022	By agreement only	US Census Bureau
Longitudinal Employer-Household Dynamics	2014 vintage	By agreement only	US Census Bureau
SSA SSR file	2010-2014, 2016, 2019-2021	By agreement only	Social Security Administration
Census Numident File	2019 Q1 vintage	By agreement only	US Census Bureau-SSA joint product
Internal Revenue Service Files			
Form W2	2005-2016	By agreement only	IRS
Form 1040	2005-2016	By agreement only	IRS
California health data			
Birth Statistical Master Files	2007-2011	By agreement only	California Department of Health Care Access and Information
Inpatient records	2007-2012	By agreement only	California Department of Health Care Access and Information
County-level US economic data			
Small Area Income and Poverty Estimates Files	2005-2016	Provided in /data/CA_data	US Census Bureau
Swedish Administrative Data			
Medical Birth Records	2007-2017	By agreement only	National Board of Health and Welfare
Death records	2007-2017	By agreement only	National Board of Health and Welfare
Inpatient records	2007-2017	By agreement only	National Board of Health and Welfare
Longitudinal Database of Individuals	2005-2016	By agreement only	Statistics Sweden
Other miscellaneous data			
Swedish CPI	2005-2016	Provided in /data/SE_data	Official Statistics of Sweden

3 Programs

All programs are divided into six groups identified by the first number of their file names,

0. Extract and build CA data
1. Clean CA data
2. Generate CA summary statistics and outcome gradients
3. Build and clean Swedish data
4. Generate Swedish summary statistics and outcome gradients
5. Generate summary statistics of previously created gradients and plot

Within the code repository, these six groups have been separated into four folders, each with its own master .do file,

1. 01_CA_Scripts
2. 02_Swedish_Scripts
3. 03_Final_Output_Scripts

Note that the Census prohibits file name information from being disclosed and both the Census prohibits direct directory paths from being shared. The paths are therefore redacted as needed in the programs.

3.1 Replication instructions

1. Edit paths in 00_master_ca_public.do and run. This will build the publicly available CA dataset, the first program in group 0 as defined above.
2. Transfer the folder 01_CA_Scripts and its contents into the Census FSRDC as well as the dataset of county-level economic estimates created in step 1.
3. Edit file paths in 00_master_ca_rdc.do and run. This will sequentially run all programs within the folder, the remaining programs in group 0 and all programs in groups 1-2 as described above.
4. Transfer the folder 02_Swedish_Scripts and its contents onto the secure server storing the Swedish administrative data.

5. Edit the file paths in 00_master_se.do and run. This will sequentially call all programs in groups 3-4 as described above.

6. Following proper disclosure and file transfer protocol of the final versions of the gradients created in steps 2 and 4, place all gradients into the appropriate sub-folder within "output/gradients".

   (a) "/chen" for the restricted Chen, Oster, Williams (2016) gradients

   (b) "/by_period" for gradients created separately for each time period

   (c) "/hospFE" for the infant mortality gradients with hospital fixed effects

   (d) "/main" for all other gradients

7. Edit file paths and run 00_master_output.do which creates and saves all tables and figures in the paper using the final CA and Swedish gradients created in master_ca.do and master_se.do.

Each program in the repository is described in Table 2 below. Table 3 follows and outlines which script creates each Table and Figure as well as the relevant intermediate outputs used.

Table 2: Detailed list of all Programs

Program	Purpose
0. Extract and build CA births data	
Run outside the FSRDC	
0A_import_SAIPE.do	Build CA county-level economic dataset from publicly available SAIPE data. (2005-2016)
Run inside the Census FSRDC	
0B_stataextracts.do	Process CA birth records, defining and harmonizing relevant variables over time. Augment parent information with Census survey data. (1995-2016)
0C_extract_mortality.sas	Extract Numident (2019, quarter 1) to generate mortality indicators.
0D_oshpd_readin.do	Read in first year of life hospital and ED records and append all years (1991-2012)

Continued on next page

Table 2 – *Continued from previous page*

Program	Purpose
0E_extract_acs.sas	Extract ACS and Census crosswalks (2005-2017)
0F_acs_readin.do	Read in ACS data, merge with Census crosswalks, and append all years (2005-2016)
0G_extract_chck.sas	Extract raw data from Census Household Composition Key (2016-2019)
0H_chck_readin.do	Clean variables in yearly CHCK data files, merging on child PIK.
1. Clean CA births data and merge with economic data (in Census FSRDC)	
1A_clean_cabirths.do	Get variables from CA births data and merge with Numident for composite death variables.
1B_extract_tax.do	Extract Federal 1040 and W2 tax data (1995-2016) and sum earnings to create family and individual income measures.
1C_extract_LEHD.sas	Extract LEHD vintage 2014 earnings data for available states.
1D_merge_births_income.do	Merge CA births sample with income measures to generate our analysis dataset of first births from 2007-2016.
1E_merge_parentdeath.do	Merge analysis dataset of births with mortality extract to get parental deaths. This saves our final analysis sample (2007-2016).
1F_identify_SMM.do	Create indicator for Severe Maternal Morbidity from OSHPD extract (2007-2012).
1G_merge_oshpd.do	Merge maternal morbidity indicator and infant mortality data with our analysis sample.
2. Generate CA summary statistics and gradients (in Census FSRDC)	

Continued on next page

Table 2 – *Continued from previous page*

Program	Purpose
2A_sumstats.do	Print first born birth rate and save the distribution of CA income in our sample.
2B_gen_gradient_main.do	Create main income gradients of outcome means.
2C_gen_gradient_chen.do	Create Chen et al (2016) gradients of outcome means.
2D_gen_gradient_hospFE.do	Create main infant mortality gradients with hospital fixed effects.
2E_gen_gradient_county.do	Create gradient means using county economic measures to construct bins.
2F_gen_gradient_main_by_period.do	Create main gradient outcome means over time.
2H_gen_gradient_main_pAGI_endpoints.do	Generate a dataset of CA income percentile endpoints (for a subsequent merge with Swedish data).
3. Build and clean Swedish births data (on restricted server)	
3A_LISAIncome_1990_2016.do	Clean income and demographic variables and adjust with CPI data.
3B_clean_MFR_Mamma_Barn.do	Clean Medical Birth Records data and store relevant variables (1995-2016).
3C_merge_MFR_parental_income.do	Merge birth records to parental controls and income.
3D_merge_MFR_pardeaths.do	Merge family deaths data to define maternal death. Save final analysis samples.
3E_merge_cA_pAGI_pct.do	Merge CA percentiles to Swedish data (created in 2H). Save the CA percentile-based analysis sample.
4. Generate Swedish summary stats and gradients (on restricted server)	

Continued on next page

Table 2 – *Continued from previous page*

Program	Purpose
4A_sumstats.do	Print parental identification rate for reference and save the distribution of income in our sample.
4B_gen_gradient_main.do	Create main gradients of outcome means.
4C_gen_gradient_main_CApct.do	Create gradients using CA income percentiles.
4D_gen_gradient_chen.do	Create Chen et al (2016) gradients by income bin.
5. Plot gradients and test significance	
5A_gradient_sumstats.do	Generate a table of outcome means for those missing income (Appendix Table B1). Plot CA and Sweden sample income distributions (Appendix Fig. C13).
5B_graph_CA_gradients.do	Plot all CA gradient figures (Figs. 1-3 and Appendix Figs. C1-C9)
5C_graph_CA_SE_gradients.do	Plot all remaining gradients with both CA and SWE distributions (Fig. 4 and Appendix Figs. C10-C12 and C14-C15).
5D_test_stat_sig.do	Test significance of outcome means by calculating p-values for reference in the text.

Table 3: Tables/Figures and Programs

Table or Figure	**Code**	**Relevant output**
Figure 1	5B_graph_CA_gradients.do	CA_birth_means_fb_2007_2016.dta
		CA_oshpd_means_fb_2007_2012.dta
		CA_inf_means_fb_2007_2011.dta
Figure 2	5B_graph_CA_gradients.do	CA_birth_means_fb_2007_2016.dta
		CA_oshpd_means_fb_2007_2012.dta
Figure 3	5B_graph_CA_gradients.do	CA_birth_means_fb_2007_2016.dta
		CA_birth_means_quintiles_by_race_fb_2007_2016.dta
		CA_oshpd_means_deciles_by_race_fb_2007_2012.dta
		CA_inf_means_deciles_by_race_fb_2007_2011.dta
Figure 4	5C_graph_CA_SE_gradients.do	CA_birth_means_fb_2007_2016.dta
		CA_oshpd_means_fb_2007_2012.dta
		CA_inf_means_fb_2007_2011.dta
		SE_gradient_ventile_means_2007_2016_fb_MFR_MB.dta
		SE_gradient_ventile_means_2007_2011_fb_MFR_MB.dta
Appendix Table B1	5A_gradient_sumstats.do	CA_birth_means_fb_2007_2016.dta
		CA_oshpd_means_fb_2007_2012.dta
		CA_inf_means_fb_2007_2011.dta
		SE_gradient_ventile_means_2007_2016_fb_MFR_MB.dta
		SE_gradient_ventile_means_2007_2011_fb_MFR_MB.dta
Appendix Table B2		HCAI and CDPH table
Appendix Figure C1	5B_graph_CA_gradients.do	CA_county_med_birth_means_fb_2007_2016.dta
		CA_county_med_oshpd_means_fb_2007_2012.dta
		CA_county_med_inf_means_fb_2007_2011.dta
Appendix Figure C2	5B_graph_CA_gradients.do	CA_county_pov_birth_means_fb_2007_2016.dta
		CA_county_pov_oshpd_means_fb_2007_2012.dta

Continued on next page

Table 3 – *Continued from previous page*

Table or Figure	**Code**	**Relevant output**
		CA_county_pov_inf_means_fb_2007_2011.dta
Appendix Figure C3	5B_graph_CA_gradients.do	CA_inf_means_fb_2007_2011.dta
Appendix Figure C4	5B_graph_CA_gradients.do	CA_birth_means_fb_2007_2016.dta
Appendix Figure C5	5B_graph_CA_gradients.do	CA_inf_means_fb_2007_2011.dta CA_inf_means_deciles_fb_2007_2011.dta hospFE/CA_inf_means_fb_2007_2011.dta hospFE/CA_inf_means_deciles_fb_2007_2011.dta
Appendix Figure C6	5B_graph_CA_gradients.do	CA_birth_means_by_race_fb_2007_2016.dta CA_oshpd_means_deciles_by_race_fb_2007_2012.dta CA_inf_means_deciles_by_race_fb_2007_2011.dta
Appendix Figure C7	5B_graph_CA_gradients.do	by_period/CA_birth_means_fb_2007_2016.dta by_period/CA_oshpd_means_fb_2007_2012.dta by_period/CA_inf_means_fb_2007_2011.dta
Appendix Figure C8	5B_graph_CA_gradients.do	by_period/CA_birth_means_by_race_fb_2007_2016.dta