

README File for Data and Programs

“What is the Added Value of Preschool for Poor Children? Long-Term and Intergenerational Impacts and Interactions with an Infant Health Intervention”

Maya Rossin-Slater and Miriam Wüst

General description of analysis strategy and data sources

For our analyses, we construct administrative register data sets on outcomes and characteristics at the individual level for cohorts born in 1930-1957, and then collapse the data to municipality x year of birth cells.

For all women born in or after 1935, we find their biological children and the children's fathers (if available in the data). Our main analysis of intergenerational outcomes focuses on the oldest children of preschool-exposed women, although we present a robustness test using all children. This data set is constructed similarly to the analysis data set for first generation outcomes, and is collapsed to the mother's municipality x year of birth level.

We merge on external data at the municipality x year level. We use the following general sources of data:

- **Population-level data on individual characteristics and outcomes from several administrative registers (provided by Statistics Denmark)**
 - We construct a data set of all individuals born in Denmark in 1930-1957 with the following variables: individual identifier, date of birth, parish of birth, and gender. We only keep individuals with a non-missing and valid parish of birth (as this allows us to assign individuals to their municipalities of birth).
 - We merge this population-level data set with data on outcomes that come from multiple administrative data sets: mortality, hospital nights and diagnoses given at hospitals; income variables; educational attainment variables.
- **External data:**
 - We use election and census data available from the Danish National Archives to add on municipality-level control variables (we interpolate between census/election years).
 - We have collected and digitized data on all government-approved preschools in Denmark from a series of publications described and referenced in the paper. We collapse these data to the municipality x year level to measure preschool exposure in every municipality x year cell (see paper for details).
 - We add aggregate data from the “Medical Reports for the Kingdom of Denmark,” which are available at the urban municipality (i.e. town) and county level, on the annual number of live births, infant, and child deaths. These data were originally collected for Wüst (2012).
 - We use data on nurse home visiting program treatment status, available at the municipality x year level for years 1937-1949. These data were originally collected for Wüst (2012) through year 1949, and we add on treatment status for post-1949 municipalities from a series of books referenced in the paper.

Access to Danish administrative register data:

The data used in our paper comes from several Danish administrative registers. The data is physically stored at Statistics Denmark and due to security considerations the data may not be transferred to computers outside Statistics Denmark. Researchers (and their assistants) are allowed to work with the data if their projects are approved by Statistics Denmark and they are affiliated with Danish research institutions. Access to the data is provided via a remote desktop.

If researchers at a university or other research institutions outside Denmark want to use Danish administrative data they may visit a Danish research institution or collaborate with researchers in Denmark. If researchers want to analyze our data for replication purposes, we will provide guidance with regard to getting project approval at Statistics Denmark, and we will provide all the programs used in our analyses.

The do-files in this directory create the analysis data from the raw register data sets and include the code used to generate all of the tables and figures in the paper and appendices. The do-files are ordered (numbering reflects rough order of execution). The majority of files is divided in two classes: labeled with “_cr_” (create a data set and clean variables) and “analysis” (predominantly for the analysis of final data). There are a few do-files that serve both purposes (and also create descriptive figures and tables).

Data build from raw external and register data:

1_cr_daycare_and_controls_data.do: Prepares external preschool and controls data at the municipality x year level.

1_cr_municontrolvars.do: Prepares external data from the Danish National Archives from elections and censuses (municipality x year level).

1_cr_parish_muni_match.do: creates a dataset that links the parish and municipal structure of Denmark.

1a_cr_underlying_pop.do: We start by using the yearly “GRUND” data sets provided by Statistics Denmark. Here each resident in the given year in DK has a personal identifier. We find all DK born individuals of the cohorts 1930-60 in these data, their date of birth and parish of birth. The do-file keeps all individuals with a “valid” parish code and a data set with a “total population” irrespective of valid parish code to compare with our analysis data.

1b_cr_pop_daycare_yearapproval.do: Merges the above population data set (1930-1957 birth cohorts) to preschool data prepared in *1_cr_daycare_and_controls_data.do*. Merges on controls (not interpolated). Output: **pop_daycare_popmunic_yrapproval.dta**

Creation of mother analysis data:

2_cr_fertility_post1935.do: Finds all mothers who are born in 1935-1960 (with valid parish of birth codes) in GRUND. Finds their biological children in FTDB2012.dta (provided by Statistics Denmark). The link is performed using the mothers’ unique personal identifiers. All mothers are then merged to the population that we can use in the analyses (have a valid parish). Merge on nurse home visiting program information, collapse the data to municipality x year

level, and add control variables. Create relevant sample of “ever preschool” municipalities. Output: **analysis_daycare_only moms_everopen.dta**

Creation of outcome datasets from raw administrative data:

3_cr_civstoutcome.do
3_cr_diagoutcomes.do
3_cr_mortality.do
3_cr_econmovingaverages.do
3_cr_econoutcomes.do
3_cr_edu_outcomes:

Create outcome data based on raw data sets provided by Statistics Denmark: Ever married; diagnoses and hospital admissions; survival/mortality; income and education of the first generation. All of these data are available for the 1980-2012 period in our data sets.

Creation of second generation analysis data:

3_cr_edu_kids.do: Finds the education outcomes for the second generation. Merges to preschool and nurse data, collapses to municipality x year level. Adds municipality-level controls. Creates the analysis dataset **analysis_daycare_intergen_everopen.dta** (for first-born children). If the do-file is run without the first-kid constraint, it provides the analysis dataset for the “all kids” analysis in the Appendix.

3_cr_kids_healthoutcomes.do: Merges kids to hospital admissions (second generation). Sample of first kids. Results not reported in paper.

Intermediate do-files that use external data at county/town level to assess effects of preschool on fertility, missing observations in our data and the first stage on enrollment:

4a_brithspercounty.do: Compares county-level number of births and survivors past age 1 to the observations in our data [APPENDIX FIGURE C1]

4b_cr_births_alltowns.do: Uses external data to create a data set at urban municipality (town) level with annual number of births

4c_analysis_towns_firststage.do: use towns data from *4b_cr_births_alltowns.do* to assess preschool enrollment (first stage) [APPENDIX TABLE A11], the impact at town level on births in years after preschool approval [APPENDIX TABLE A1], and the “number of missing observations” [APPENDIX TABLE C1]

Creation of first generation main analysis dataset:

5_cr_analysis_munic_yearapproval.do: uses **pop_daycare_popmunic_yrapproval.dta** and merges on all outcome data sets created in “3”. Merges on nurse home visiting program treatment status for full sample. Calculates summary statistics [TABLE 1]. Then constrains the data set to “ever implementing” municipalities. Creates human capital index outcome variable. Regresses all outcomes on month of birth and gender fixed effects to calculate conditional mean outcomes in municipality x year cells, take residuals. Collapses the data to municipality x year cells. Adds in municipality x year controls.

- Option in this do-file: omit “bottom p1” and create analysis sample for the robustness test [APPENDIX TABLE A14]

Analyses do-files and their output:

6a_analysis_munic_yrapproval_descr.do:

TABLE 2: correlation of nurse home visiting (NHV) program and preschool treatment

TABLE 3: observables and treatment initiation

APPENDIX FIGURE A1: Variation in NHV program and preschool

APPENDIX FIGURE A2: Difference in years between in NHV and preschool treatment

APPENDIX TABLE A9: do predicted outcomes correlate with treatment?

FIGURE 1B: Share of municipalities with approved preschool by year

6b_analysis_munic_yrapproval_eventrgraphs_age11.do:

FIGURE 2 and APPENDIX FIGURE A5: event-study graphs, first generation

6c_analysis_munic_yrapproval_regs.do: creates main results and robustness, first generation

TABLE 3: main results

TABLE 6: interaction results

APPENDIX TABLE A2: alternative income measures

APPENDIX TABLE A3: hospitalization outcomes

APPENDIX TABLES A7 AND A8: robustness to chance of specification or sample

APPENDIX TABLE A12: interaction result robustness

APPENDIX TABLES B1, B2, B3, B4: main effect of NHV program, main effect of NHV and preschool (no interaction)

6d_analysis_momfertility.do:

APPENDIX TABLE A5: effects of preschool on fertility of first generation

6e_analysis_mainout_onlymoms.do: main results for sample of moms only (collapsed data).
Not reported as table in paper.

6f_analysis_yrapproval_p1_hc.do:

APPENDIX TABLE A14: Robustness to omitting bottom p1

6g_permut.do and *permut.ado:* APPENDIX FIGURE A5

7_analysis_intergenerational.do: analysis for second generation

TABLE 5: main outcomes for second generation (first-born kids)

APPENDIX TABLES A7 AND A8: robustness to specification and sample

FIGURES 2 AND A4: event-study graphs for second generation

APPENDIX TABLE A10: predicted outcomes

Health effects for second generation and interaction results (not reported in paper)

7_analysis_intergenerational_allkids.do:

APPENDIX TABLE A6: robustness to using all kids in the second generation

8_analysis_individual_level.do: APPENDIX TABLE A4: individual-level regression results

8_intergen_transmission_coef.do: Intergenerational transmission of education. Reported in the paper text but not in table format.

Additional do-files:

Map.do: creates FIGURE 1A (map of DK)

First.do: assigns directories and globals

Tabbottom.do: table bottom formatting