README file for Conner, Einav, Finkelstein, and Persson "Targeting Precision Medicine: Evidence from Prenatal Screening"

Overview

The code in this replication package describes the data and programs used in the paper referenced above.

Data Availability and Provenance Statements

Our main data described in Section 2 of the paper is sourced from Statistics Sweden, the Pregnancy Register, and the National Board of Health and Welfare. The data contain sensitive information on individual outcomes and hence, Swedish law requires users of the data to hold a permission from the Swedish Ethical Review Authority (ERA, "Etikprövningsmyndigheten"). This implies that we are unable to make the data directly available online. Researchers interested in obtaining the data could apply for a permission from the ERA at https://etikprovningsmyndigheten.se/. The cost of applying for the permission is currently 16,000 SEK (about 1470 dollars as of April 2024), and the processing time is approximately 3 months. After having obtained an ERA approval, the next step is to order the data from Statistics Sweden, The Pregnancy Register, and the National Board of Health and Welfare (we are not able to provide information on the precise costs involved). In this process, Statistics Sweden assigns each individual an identification number unrelated to the actual personal identification number used in Sweden. This identifier then makes it possible to match data from the different sources/administrative registers described above (the identifier is sent by Statistics Sweden to the National Board of Health and Welfare and the Pregnancy Register in order to obtain information from these registers). Statistics Sweden, the National Board of Health and Welfare, and the Pregnancy Register carry out all these steps. Thus, the datasets that are delivered to the researcher do not contain personal information such as the (true) personal identification number, or the name and address. The data will be 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 EU). All handling of the data must satisfy all applicable Swedish laws.

All program code is made available and described in detail below.

Computational requirements

Software Requirements

- Stata (code was last run with version 17.0)
- R (version 3.6.3) and RStudio (version 1.2.5033)

Description of programs/code and Instructions to Replicators

The provided code reproduces all tables and figures in the paper and appendix.

Programs

Files in the */derived/* folder generate the various samples we reference in the paper, including the baseline sample and the full sample. The code in this folder is split into the following subfolders:

- master.do: runs the same construction scripts (excluding */nt_data_coverage*) in the correct order.
- */integrate_kub_data*
    - build.do: cleans the updated KUB data and merges in the original KUB data; fomr these, it constructs testing, testing outcome, and pregnancy outcome variables.
- */make_master_dataset*
    - build_new_mfr.do: cleans the birth registry data and the KUB data from the previous step. It calculates birth dates, adds birth diagnoses, and adds test ids. It also matches tests to pregnancies to create a combined master dataset. Applies the following sample restrictions:
        - Drop non-singleton pregnancies/tests
        - Drop if missing NT risk scores
        - Drop tests before 2/1/2009
    - clean_new_mfr.do: cleans the above master dataset by defining dates, counties, age, testing variables, outcomes, CA, fetal issues. The end product is the cleaned master dataset.
- */make_waves*
    - build.do: creates a dataset called “waves” that defines when cfDNA coverage is introduced for each county.
- */build_income_tiles*
    - build.do: uses tax data and the birth registry data to calculate income quartiles and ranks for mothers
- */make_analysis_sample*
    - build.do: builds the analysis sample by applying sample restrictions to the cleaned master datasets. Drops observations with test years outside of 2011-2019, and keeps only pregnancies in universal NT coverage regions. Also merges waves and mother characteristics to the sample and generates useful variables. The end product is the full sample.
- */build_model_sample*
    - build_v2.do: creates preliminary baseline sample from the analysis sample and the birth registry data and dropping duplicates. It also adds relevant covariates such as mother characteristics. The end product is an intermediate dataset.
    - clean_sample.do: cleans the intermediate dataset by generating useful variables. Creates an all live births sample, an all live births in universal NT region months

sample, an all NT screens in universal NT region-months sample, and the baseline sample. The latter is created by limiting the all NT screens in universal NT region-months sample to those with NT score greater than or equal to 1/1000.
    - analysis.do: creates a table of summary statistics of each dataset created in clean_sample.do. The end product is Appendix Table A3.
- */nt_data_coverage*
    - analysis.do: calculates take-up of NT in universal NT coverage regions.

Files in the */descript_by_xs/* folder generates Figure 2 of the paper (Post-NT testing rates by demographics).

- analysis.do: calculates post-NT testing rates of baseline and full samples by demographics. The conditional rates are calculated by regressing the rates on indicators for NT risk score bins, reporting the average residual plus the sample mean.

Files in the */draft_tab_fig/* folder generate several exhibits in the paper (Figure 1, Figure 3, Table 1, Appendix Figure A1, Appendix Figure A2, Appendix Figure A3, Appendix Figure A4, Appendix Table A1, Appendix Table A2).

- descript_tables.do: calculates summary statistics and generates Table 1. It calculates the share of pregnancies in each NT coverage regime (Appendix Table A2) and cfDNA coverage regime (Appendix Table A1).
- rf_figures.do: calculates risk score distributions and testing rates of different samples and coverage regimes. It outputs Figure 1, Figure 3, Appendix Figure A1, Appendix Figure A2, Appendix Figure A3, and Appendix Figure A4.
- rf_tables.do: examines outcome shares by risk level and testing differences between waves.

Files in the */estimate_w_xs* folder estimate the main model and generate some model-related exhibits (Figure 5, Table 2, Table 3, Table 4, Appendix Figure A5, Appendix Figure A6, Appendix Figure A7, Appendix Figure A8, Appendix Table A6, Appendix Table A7).

- master_run.do: runs the post-estimation scripts in order.
- estimate_xs.R: estimates the structural model via method of moments, matching cfDNA and invasive testing decisions for risk score and coverage regime bins. The model parameters are outputted in the log file and are in the paper as Table 2.
- production_model_xs_new.R: R script that defines functions used in estimation-related R scripts in */estimate_w_xs/* and */results_exhibits_xs/.*
- counterfactuals_xs.R: calculates counterfactual testing decisions, outcomes, and welfare under different coverage regimes and information/technology availability. The results are outputted in the log file and are in the paper as Table 3 and Table 4.
- counterfactuals_xs_full.R: calculates counterfactual testing decisions, outcomes, and welfare under different coverage regimes and information/technology availability for

the full sample. The results are outputted in the log file and are in the paper as Appendix Table A6 and Appendix Table A7.

- counterfactual_functions.R: R script that defines the functions used in counterfactuals-related R scripts in */estimate_w_xs/* and */results_exhibits_xs/.*
- bootstrap_se.R: calculates bootstrapped standard errors that are added to Table 2.
- model_fit.do: creates model fit plots that compare the actual data moments with the model moments. These are in the paper as Figure 5, Appendix Figure A5, Appendix Figure A6, Appendix Figure A7, and Appendix Figure A8.

Files in the */results_exhibits_xs* folder produce post-estimation figures from counterfactual calculations and comparative statics (Figure 6, Figure 7, Figure 8, and Figure 9).

- model_comp_stat_xs.R: after the estimation, creates dataset used in comparative static simulated decisions and calculations. It calculates willingness to pay and predicted testing decisions under different percentiles of *a* and *c.*
- cf_figures_xs.do: produces counterfactual-related figures and comparative statics figures. These are in the paper as Figure 6, Figure 7, Figure 8, and Figure 9.

Files in the */robustness* folder run the full bootstrap for the welfare numbers (Appendix Table A4 and Appendix Table A5) and estimate variants of the model to test model robustness (Appendix Table A9).

Additionally, some numbers in the text are found by tabulating from samples or counterfactual calculation tables:

- *Share of singletons:* found in "tab nonsingleton" of */derived/make_master_dataset/build_new_mfr.do*
- *Take-up of NT in universal NT regions:* the calculation procedure is described in Appendix A2; the code is found in */derived/nt_data_coverage/analysis.do*
- *Unique women in full sample:* outputted at "unique lopnr" of */derived/build_model_sample/analysis.do* with `samp' = 3
- *Unique women in baseline sample:* outputted at "unique lopnr" of */derived/build_model_sample/analysis.do* with `samp' = 4
- *Pregnancies in our baseline sample cover X% of full sample CA pregnancies:* calculation follows from footnote 24 in the text; the summations come from the summing over risk scores in the baseline and full samples.
- *Number of pregnancies in baseline sample that result in live birth with CA:* count if "chroma = 1" in baseline sample.
- *Share of draws with $c_i > a_i$:* outputted in counterfactuals_xs.R log file
- *Share of patients with high enough NT risk score that they would want an invasive test even after negative cfDNA:* Table 3, "Invasive only" row of "Free invasive & Full cfDNA covg. column"

- *Share of patients that would avoid invasive testing even after positive cfDNA result:* calculated from Table 3 as {100 - "cfDNA only" row of column 4 - *Share of draws with* $c_i > a_i$ - "Invasive only" row of column 4}

Figure 4 and Appendix Figure A9 are not generated in the code/data.

Notes on helper functions and codebooks

There are many Stata packages and R libraries that are used in the code. These are included in the */ado* (for Stata) and */R-library* (for R) folders. Furthermore, much of our code is based on relative paths, so it is necessary to define global file paths at the beginning of code files.