# "Adverse Selection and an Individual Mandate: When Theory Meets Practice"

Martin Hackmann, Jonathan Kolstad, and Amanda Kowalski

Guidebook for the Empirical Analysis

This file describes the key steps of the empirical analysis and is organized into four sections. The first section provides a brief overview of the directory structure used in this analysis and introduces the master file that calls different programs and conducts the entire empirical analysis. The second section discusses the preparation of the raw proprietary SNL and NHIS data. The third section discusses continuing data cleaning efforts for various data sets, which are ultimately used in the main analysis. We discuss the main analysis in the last section, which produces the empirical results that are presented in the paper.

## 1) Directory Structure , Master File and Package

The directory structure for the databases and stata do-files is as follows:

- The do files are saved in the chosen local directory "localdir", except for the data preparation code for the NHIS data, "Prepare_NHIS.do" which we conducted at the Census Data Center. This do file is located in "localdir\Data\NHIS". We will come back to this analysis in the next section.
- The program assumes that the raw data sets are stored in the following directories
    - The proprietary raw SNL data are located at "localdir\Data\SNL"
    - The proprietary NHIS data are located at "localdir\Data\NHIS"
    - The data from the American Community Survey are saved at "localdir\Data\ACS"
    - The data from the Medical Expenditure Panel Survey are saved at "localdir\Data\MEPS"

  We will comment again on the location of these datasets in the following sections.
- The complied databases, compiled in the analysis described in section three, built on these raw data and are saved in "localdir\Data\Compiled"

Unix/Linux users will have to change the back slashes "\" into forward slashes "/". Unfortunately, these slashes cannot be changed globally as it may cause errors in included matrix commands (e.g. matrix M=v1 \ v2) as well as in codes that produce LaTeX files.

The master file, "Master.do", is located in "localdir". To replicate the empirical analysis, the researcher will have to adjust the local directory, defined in this master file accordingly. Notice however, that various data sets are proprietary and cannot been shared. For details, see the next sections.

Synth package: The synthetic control method uses the Stata package "synth", which may not be installed yet. Please install this package prior to implanting the code.

**2) Raw Data Preparation:**

SNL data:

The do-file "Prep_snl_master.do" calls the relevant do-files "Prep_snl1.do", "Prep_snl2.do", "Prep_snl3.do", and "Prep_snl4.do", which create the intermediate dataset "datasnl.dta" from raw SNL excel files. These data are proprietary and cannot be shared. The code assumes that the data are located in the directory "localdir\Data\SNL". The program produces an intermediate data set called "datasnl.dta", which is saved in to the "localdir\Data\Compiled" directory.

NHIS data:

Prepare Raw data:

- Prior to our Census data center visit, we combined the NHIS family and person-level data for the years 2004-2010. We also added information on the Medical Consumer Price Index. Our host at the Census data Center merged these databases with a state identifier and saved the final database under the name "data.dta".
  The most important variables are:
  Age_p: to identify the relevant age group
  Medicare, Medicaid, schip, his, military, otherpub, othergov: to identify health insurance coverage
  Hiempof: to measure whether the individual was offered health insurance through the employer
  Rat_cat: to quantify the income of the individual
  Wtfa: to weight the observations
  Stratum, strat_p:
- Next, we ran the do-file "Prepare_NHIS.do" at the Census Data Center, which calculated the key databases "state-period-mean.dta" and "observations-all.dta" , which we use in the continuing data analysis, see the prepare data section. These data are proprietary and cannot be shared. The do-file "Prepare_NHIS.do" is located in "localdir\Data\NHIS". The codes assume that the databases are located in the same directory.

**3) Prepare Data:**

The do-file "Preparedata.do" loads and cleans the following datasets:

- "state-period-mean.csv" and "observations-all.csv"
  These are NHIS data, accessible only through Census data centers, provide information on post-reform coverage estimates by state as well as the underlying number of observations. Combining this information we can construct the standard deviation of the coverage estimates, which are used in the bootstrap analysis. These intermediate data sets were generated from

raw NHIS data, see the raw data preparation section for details. This dataset is proprietary and cannot be shared. The code assumes that the data is located in the directory “localdir\Data \NHIS”

- “State-Match.csv”
  This database links the state name with the state id from the former two datasets. This dataset is proprietary and cannot be shared. The code assumes that it is located in n the directory “localdir\Data\NHIS”
- “medcpi04-11.csv”
  This dataset contains the Medical consumer price indices for the years 2004 through 2011, which are normalized to 2012. These indices are used to normalize premium and claim expenditure information to 2012 dollars. The code assumes that this file is located in the directory “localdir\Data”
- “hmo-info-plan.csv”
  This dataset contains two variables, the name of the insurance carrier and an indicator whether the name of the carrier has the HMO in it. This information is used in a robustness check, see the paper for details. This dataset is proprietary and cannot be shared. The code assumes that it is located in the directory “localdir\Data\SNL”
- “datasnl.dta”
  This is the main database for the analysis. It contains the key variables of interest at the carrier-state-year level. The dataset was constructed from SNL raw data, see the raw data section for details. This dataset is proprietary and cannot be shared. The code assumes that it is located in the directory “localdir\Data\SNL”

The key final product of this do-file is the data set “datasnl-final.dta” which joins and cleans the underlying datasets and aggregates the information to the state-year level.
The key variables of interest for the co*n*tinuing analysis are:

- “prem_per_year_ind”:
  This variable measures the average annual premium at the state-year level in the respective individual market.
- “prov_health_serv_incur_ind_year”:
  This variable measures the average annual claim expenditures per enrollee at the state-year level in the respective individual market.
- “Mem_mo_ind”:
  This variable measures the total number of enrolled people at the state-year level in the respective individual market.
- “Year”:
  This variable denotes the respective year
- “state”:
  This variable identifies the respective state
- “insurance300post”:

This variable measures the percentage of enrolled people in the post-reform years (2008-2010) in the respective state. Specifically, this is the fraction of insured people amongst people who earn more than 300% of the federal poverty line who are neither eligible for premium subsidies nor for Medicaid. This variable was created using the NHIS data and is used to express observed enrollment in the SNL data in percentages.

There are two additional databases that are used in the robustness check analysis:

- "dataacs.dta"
  This database contains demographic information from the American Community Survey (ACS) at the state-year level. This information is used in our robustness analysis with respect to contemporary demographic trends, see the description of "Demo_robcheck_pointestimates.do" below.
- "datameps.dta"
  This database contains health spending information on emergency room visits, inpatient and outpatient services, spending on prescription drugs, as well as other expenditures from the Medical Expenditure Panel Survey (MEPS) . This information is used in the robustness check analysis, where we revisit our assumption of perfect community rating. The key variables here are:
  - "agex" which describes the age of the individual
  - "allxp" which is the sum over the health care expenditures on the services listed above

**4) Main Analysis:**

Here, we conduct the main analysis of the paper. The do-file "Main_analysis.do" calls various do-files that subsequently produce the results presented in the paper. This do-file exploits the data sources mentioned in the previous section and calls the following do-files:

a. "Abadieweights.do"
   This do-file calculates the synthetic control weights used in the main analysis. We construct two sets of weights, weights for the main analysis and weights that emphasize the role of guaranteed issue states. The latter weights are used in the robustness check analysis. The weights are saved the database "abadieweights.dta"

b. PointestimatesDID.do
   In this do-file, we calculate the point estimates for the difference-in-differences regressions tables. There are two tables, the table representing the baseline regression results as well as the regression table that emphasizes the role of the guaranteed issue states. The baseline and the guaranteed-issue point estimates are saved in "pointestbaseline.dta" and "pointestgi.dta" respectively.

c. Bootstrap_DID.do

In this do-file, we calculate the confidence intervals for our baseline regression table. We use a block bootstrap approach and save the results in the table "did-table-boot.tex". We also calculated alternative confidence intervals based on an estimate of the standard deviation (instead of calculating) the $2.5^{th}$ and the $97.5^{th}$ percentile. The results are almost identical and saved in the table "did-table-boot-sd.tex". We present the confidence intervals for our key parameters of interest in the online appendix.

d. Bootstrap_DIDgi.do
In this do-file, we calculate the confidence intervals for our regression table that emphasizes the role of guaranteed issue states. We use a block bootstrap approach and save the results in the table "did-table-boot-gi.tex". We also calculated alternative confidence intervals based on an estimate of the standard deviation (instead of calculating) the $2.5^{th}$ and the $97.5^{th}$ percentile. The results are almost identical and saved in the table "did-table-boot-gi-sd.tex". We present the confidence intervals for our key parameters of interest in the online appendix.

e. Bootstrap_welfare.do
In this do-file, we calculate the mean estimates and the confidence intervals for our welfare effects. We use a block bootstrap approach and save the results in the csv file "bootstrap-welfare.csv". The table is organized as follow. Each row corresponds to a different tax penalty as shown in the paper. The last row shows the effect for the baseline penalty but using the weights that emphasize the role of guaranteed issue states. The first column shows the full welfare effect followed by a column that indicates the statistical significance and two columns that show the respective confidence intervals. Column 5 indicates the fraction of simulated full welfare effects that are greater than 0. This column is, however, not reported in the paper. The columns 6-9 show the analogous results for the net welfare effects. Again column 10 indicates the fraction of simulated net welfare effects that are greater than 0. This column is, however, not reported in the paper.
Finally, we use an excel macro, which converts this information to the table presented in the paper.

f. DIDgraphs.do
In this do-file, we construct the baseline difference-in-differences graphs for coverage, claim expenditures, and premiums.

g. DIDgraphsgi.do
In this do-file, we construct the difference-in-differences graphs that emphasize the role of guaranteed issue states for coverage, claim expenditures, and premiums.

We conduct the robustness checks in the following do-files, which are also called from the aforementioned do-file "Main_analysis.do":

h. Demo_robcheck_pointestimates.do
In this do-file, we construct the point estimates for the robustness check table that controls for demographic trends. We load additional demographic information for the years 2005-2010 from the American Community Survey (ACS) contained in the database "dataacs.dta". Please contact the authors for more information on the raw data and the programming that led to this summary database. We save the point estimates in "pointestrob.dta".

i. Demo_robcheck_bootstrap.do
In this do-file, we calculate the confidence intervals for the robustness check table that addresses contemporary demographic trends. We use a block bootstrap approach and save the results in the table "did-table-boot-demorob.tex".

j. MEPS_robcheck.do
In this do-file, we quantify the statistical relationship between age and health care utilization, which we cite in the robustness check analysis regarding our community rating assumption. To conduct this exercise, we load additional data on health care utilization from the Medical Expenditure Panel Survey (MEPS) for the years 2004-2010, contained in the database" datameps.dta". We combine spending information on emergency room visits, inpatient and outpatient services, spending on prescription drugs, as well as other expenditures. Please contact the authors for additional information on the raw data.

k. hmo_robcheck.do
In this do-file, we test the robustness of our findings with respect to potential changes in plan generosity. To this end, we use the carrier name information into HMO and Non-HMO plans. We present the regression results for premiums and average costs that control for the share of individuals that are enrolled in HMO plans. The point estimates are very similar to our baseline estimates.

l. placebo_master.do
In this do-file, we conduct the placebo test outlined in the online appendix. This do-file loops over placebo states and calls two do-files, "placebo_weights.do" and "placebo_welfare.do". In the former do-file, we construct the synthetic control weights assuming that the reform took place in the respective placebo state. In "placebo_welfare.do", we then calculate the respective time series for our key dependent variables which represent the deviations in the placebo treatment state from the placebo controls in each year. The "placebo_master.do" file finally combines the time series, constructs the mean squared prediction error, and highlights the relevant time series from the paper in the data browser. We copied these time series into excel and then produced the figures from the paper.

m. Alternativetobootstrap.do

In this do-file, we conduct an alternative, more parametric, approach to specify the confidence intervals in our main regression tables, Table 2 and Table 4. We compare the confidence intervals to those derived from two block bootstrap approaches conducted in “Bootstrap_DID.do” and “Bootstrap_DIDgi.do”, respectively. We discuss the details in the online appendix. The results are saved in “did-table-boot_extension.tex”.