This readme file lists and describes the programs and data used to generate the results appearing in the published version of "Blue Collar Booms and American Mortality: Evidence from the Fracking Revolution”

Data Availability

Summary

- ☐ All data are publicly available
- ☑ Some data cannot be made publicly available
- ☐ No data can be made publicly available

This replication package uses both publicly available and restricted-access data sources. See the Data Sources section below for specific access requirements for each dataset.

Software:

The code was last run on a 2-core Intel-based laptop with Windows 10 Home Edition version 22H2 with 16GB of RAM. Time estimations below are based on this.

The files were run using Stata SE Version 16.0. The code uses frames (introduced in Stata 16) to create figures. Users may have to install the following packages from the Boston College Statistical Software Components (SSC) before running the code:

```
ssc install reghdfe
ssc install ftools
ssc install coefplot
ssc install outreg2
ssc install spmap
ssc install shp2dta
ssc install ppmlhdfe
```

See https://www.stata.com/support/ssc-installation/ for instructions of how to install commands from SSC.

Data Sources

Specific and detailed explanation of the data sources used are provided in the text of the paper. Supplemental information about each dataset is provided here.

Publicly Available Data

- EIA Shale Play Shapefiles
- QWI / LEHD employment data
- SAHIE health insurance data
- SEER Population Estimates
- USDA Snap Participation Data
- Fatality Analysis Reporting System (FARS)

Restricted Access Data

- **NVSS Mortality Files**: CDC application and approval required
- **Enverus Well Production**: Academic data agreement required
- **Rystad Energy NASMaps**: License agreement/purchase required

1. Fracking Data

- **Shale Play Boundaries**: U.S. Energy Information Administration (EIA) shapefiles
    - Download: EIA Shapefiles – Shale Play Boundaries
- **Well Production**: Enverus (DrillingInfo) well-level production database (academic license required)
    - **Access:** University Partnership at Enverus. Contact Enverus's University outreach team to access data.
        - Note: We last retrieved data through their academic outreach program in late 2019, so the process by which one receives the data may have changed.
- **Prospectivity Index (RPI)**: Rystad Energy NASMaps data product
    - Note: We purchased GIS Prospectivity Data Shapefiles directly from Rystad Energy in November 2020. The process by which one receives the data may have changed; users can contact Rystad about acquiring data through an academic license.
    - We only include the county level aggregates (which are used in the main analysis) in accordance with our data agreements with them.

2. Employment and Earnings Data

- **Source**: U.S. Census Bureau, Longitudinal Employer–Household Dynamics (LEHD) program
- **Dataset**: Quarterly Workforce Indicators (QWI)
    - Download: QWI Explorer or LEHD Public Use Data

3. Mortality Data

- **Dataset**: Restricted-use National Vital Statistics System (NVSS) mortality files, 1990–2018
    - **Access**: Application required at CDC NVSS Restricted Data. See 'How to Apply' section for instructions and required forms.
- **Population Denominators**: National Cancer Institute's SEER program county-level population estimates
    - Download: SEER Population Data

4. Health Insurance Data

- **Dataset**: Small Area Health Insurance Estimates (SAHIE)
    - Download: SAHIE Data

- **Dataset**: US Department of Agriculture (USDA) SNAP Data System
    - Download: SNAP Program Data

5. Vehicle and Truck Accident Data

- **Dataset:** Fatality Analysis Reporting System (FARS)
    - **Download:** FARS Data

Instructions to run code:

The replication uses the following file structure:

|--Master Directory
    |--do
    |--dta
    |--out
        |--appendix

Researchers should download these directories to a master directory location of their choice, which we refer to as the root directory ($root) here. **Within this root directory, users must create a folder entitled 'out', and within 'out' a subfolder entitled 'appendix'.** In the do-file "Master.do", users should change the root file to their specified directory. For instance, the code

glo root "C:\Users\vmoorthy\Documents\Replication Package"

in the beginning lines of the Master.do file should be changed to the user's chosen master directory.

The package contains 7 do files, including the "Master.do" file which calls all other do files.The main results in the analysis are ultimately saved in the "$root/out" subdirectory, and appendix files in the "$root/out/appendix" subdirectory.

The do files are as follows:

1. **Master.do:** Stata do file that calls all the files for reproducing the main results. This do file is sufficient to reproduce all tables and figures in the text. Running the master file runs every do file provided.
    - Run time: Main tables and figures: 5 minutes. Appendix Figures and Tables: 50 minutes.
    - All .do files are called through **Master.do**. If you wish to run *all* results, simply run **Master.do** in full.
    - **Note:** Many results require restricted access data. Without these variables, the user will run into an error for 'variable not found'.

        i. For instance, tables 1 and 2, and figures 1b, 3, 4 and 5 in the main text utilize restricted access oil and gas and/or mortality data.
    - If you wish to run *only one result* (e.g., a single table or figure), you cannot just highlight a portion of the code in the following do files and run it in isolation. Doing so will cause errors because macros and other definitions from the preamble in Master.do will be missing.**Example: Running Table 1 only**
        i. First, run the preamble in **Master.do** (this defines macros and global options).
        ii. Then, in **Tables.do**, highlight and run **all lines corresponding to Table 1**.
        iii. For Table 1, this means selecting lines **7 through 301 (inclusive)**.

2. **Tables.do:** Stata do file that creates Tables 1 - 3 in the main text.
3. **Figures.do:** Stata do file that creates Figures 1 - 5 in the main text.
4. **Appendix Tables.do:** Stata do file that creates Online Appendix Tables B.1-B.10.
5. **Appendix Figures.do:** Stata do file that creates Online Figures A1-A.22.
6. **eventstudy_program.do** and **eventstudy_statespec.do**: Author created programs that generate the event study graphs used in paper.

Several tables are exported to Latex, and then formatted separately in Latex to the form produced in the manuscript. These changes only affect the formatting of the tables. Further, several tables and figures contain multiple panels. The do files above often export each panel as a separate document, which then can be manually combined in order to recreate the formatting structure in the paper.

The Datasets in the dta folder are as follows:

1. **final_playonly.dta:** County year level main dataset used for most results in draft. Includes only states that reside over shale plays. All restricted data mentioned above are removed from this dataset.
2. **cntyfips2000coords.dta** and **tquartilemap.dta:** shape file datasets converted to .dta files for maps.
3. **fracking_production.dta:** County-year level oil and gas production data. All restricted data is removed from this dataset.
4. **allstates.dta**: Main dataset, including all states (i.e., even those not residing over a shale play). All restricted data mentioned above are removed from this dataset.
5. **SAHIE.dta:** County year level health insurance data from the Small Area Health Insurance Estimates.
6. **SNAP.dta:** County year level SNAP participation data from the USDA.

Data Notes

This project uses restricted data that cannot be shared directly. However, we endeavored to make the variable naming convention and labeling system fully transparent, so that researchers

with access to the same data can recreate the variables needed to run the code. Once a user obtains raw restricted data, the authors are willing to assist with reasonable requests to create the final variables used in the analysis.

Mortality Data Codebook and Naming Convention

Mortality variables are composed from five parts:
1) prefix 2) cause-of-death code 3) gender 4) ethnicity 5) age group (if applicable)

Prefixes and patterns:
Crude age-specific death rates (per 100k population at that age):

- cdrdcX_genY_ethZZ_aAABB

Raw counts (number of deaths, age-specific):

- ndcdcX_genY_ethZZ_aAABB

Age-adjusted death rates (using 2000 as the standard population year; no age suffix because they aggregate over all ages):

- adr2000dcX_genY_ethZZ

Where:

- dcX = Cause-of-death category (see full list below).
- genY = Gender indicator:
  - gen0 = All genders
  - gen1 = Men
  - gen2 = Women
- ethZZ = Ethnicity indicator (if applicable). In the examples below, 00 represents the pooled sample (all ethnicities).
- aAABB = Age group indicator, e.g.:
  - a0024 = ages 0–24
  - a2544 = ages 25–44
  - a4564 = ages 45–64
  - a6599 = ages 65–99

Example:

1. cdrdc1_gen1_eth00_a2564
   - cdrdc1 = crude death rate for suicide deaths (dc1)
   - gen1 = Men
   - eth00 = All ethnicities
   - a2564 = Ages 25–64

   This variable therefore represents suicide deaths among men, ages 25–64.

2. ndcdc1_gen1_eth00_a2564
   - ndcdc1 = raw suicide death counts (dc1)
   - gen1 = men

- eth00 = all ethnicities
- a2564 = ages 25–64

Meaning: number of suicide deaths among men ages 25–64.

3. adr2000dc1_gen0_eth00
   - adr2000dc1 = age adjusted death rate (using 2000 population) for suicides (dc1)
   - gen0 = all genders
   - eth00 = all ethnicities

   This variable is therefore the age-adjusted death rate (standardized to year-2000 population) for suicides (dc1) (Note: no age suffix because the age adjustment is computed from all ages.)

Cause-of-Death Codes (dcX)

Code	Description
dc0	All-cause deaths
dc1	Suicide deaths
dc2	Alcohol-related deaths
dc3	Drug-related deaths
dc4	Cardiovascular deaths
dc5	Cancer deaths
dc6	Respiratory deaths
dc7	Infection-related deaths
dc8	Brain disease deaths
dc9	Kidney/urethra deaths
dc10	Nutrition-related deaths
dc11	Motor vehicle deaths
dc12	Other accident deaths
dc13	Homicide deaths
dc14	Pregnancy deaths
dc15	Perinatal deaths
dc16	Congenital deaths
dc17	Not classified deaths
dc18	Infection deaths
dc19	Neoplasm deaths
dc20	Blood deaths
dc21	Endocrine deaths
dc22	Mental disorders deaths
dc23	Nervous system deaths
dc24	Circulatory deaths
dc25	Respiratory deaths
dc26	Digestive tract deaths

dc27 | Skin deaths
dc28 | Skeletal deaths
dc29 | Genital/urinary deaths
dc30 | All internal deaths
dc31 | All external deaths
dc32 | Internal deaths (excluding cardiovascular)

Each of the base components (dcX, genY, ethZZ, aAABB) is included in the dataset with variable labels. For example, the variable dc23 is missing for all values in the dataset but carries the label “Nervous system deaths.” If you run ‘describe dc23’ in Stata, you will see that label. Likewise, the gender variables are labeled:

- gen0 → All genders
- gen1 → Men
- gen2 → Women

Age categories and ethnicity codes are also provided in the dataset with descriptive labels.

These labels provide a built-in codebook that makes it straightforward to verify the components of any constructed variable.

Population variables follow a parallel convention:

- pop_genY_ethZZ_aAABB
  Example: pop_gen1_eth00_a2564 = male population ages 25–64 (pooled ethnicity).

<u>Fracking well level oil and gas production data</u>

Allprod_hwell is the total value of oil and gas production from fracking wells by country. In the dataset fracking_production.dta, Allprod_hwell0 and allprod_hwell1 are the total value of oil and gas production from fracking counties in control and treatment counties, respectively. Once users obtain oil and gas production variables from Enverus, these variables can be constructed by summing the total amount of oil and gas production from all non-vertical wells (horizontal, directional, and unconventional wells).