Replication Kit Overview: "Lives vs. Livelihoods: The Impact of the Great Recession on Mortality and Welfare"

April 11, 2025

Abstract

This guide provides a comprehensive review of the data pipeline of "Lives vs. Livelihoods: The Impact of the Great Recession on Mortality and Welfare." We share the broad pipeline structure for this project, each of the data sources employed, how these data are cleaned and analyzed, and ultimately how output corresponds to exhibits in the final published paper.

Contents

1 Introduction **2**

2 NCHS pipeline **3**

2.1 Data Sources: "raw_data" folder . . . 3

2.1.1 Overview of most commonly used data by category . . . 3

2.1.2 Inventory of raw data folders . . . 3

2.2 Data cleaning: "build" folder . . . 6

2.2.1 Written descriptions of scripts . . . 6

2.2.2 Relevant figures and tables . . . 7

2.3 "Analysis" folders . . . 8

2.3.1 Main mortality event studies . . . 9

2.3.2 Sensitivity analyses . . . 9

2.3.3 Mechanisms: . . . 10

2.3.4 Aggregation / descriptive: . . . 11

2.4 "Output" folders . . . 11

3 Medicare pipeline **11**

3.1 Data Sources: "raw_data" and "restricted_data" folders . . . 11

3.1.1 Overview of most commonly used data . . . 11

3.1.2 Inventory of raw data folders . . . 12

3.2 Data cleaning: "build" folder . . . 13

3.2.1 Written descriptions of scripts . . . 13

3.2.2 Relevant figures and tables . . . 14

3.3 Data analysis: "analysis" folder . . . 14

3.3.1 Mortality event studies, overall and by subgroup: . . . 15
3.3.2 Sensitivity analyses: . . . 15
3.3.3 Mechanisms: . . . 15
3.3.4 Aggregation / descriptive: . . . 15
3.3.5 Relevant figures and tables . . . 16
3.4 "Output" folder . . . 16

4 Welfare pipeline **16**
4.1 Data . . . 16
4.1.1 Raw Data . . . 16
4.1.2 Cleaned Data . . . 17
4.2 Code . . . 17
4.2.1 Reports . . . 17
4.2.2 Scripts . . . 17

5 HRS pipeline **17**
5.1 Data . . . 18
5.2 Analysis . . . 18

6 Appendix **18**
6.1 Acronyms . . . 18

1 Introduction

This resource is a guide for how the pipeline operates for the project "Lives vs Livelihoods: The Impact of the Great Recession on Mortality and Welfare," from raw data to paper outputs.

Note that, for this project, all analysis is not conducted within a single data pipeline. Rather, it is dispersed across four discrete locations. This is a function primarily of data restrictions. This document is organized into four segments for these four data sources, namely:

1. NCHS (Section 2): The bulk of our work uses National Center for Health Statistics (NCHS) mortality data. In this portion of the pipeline, we investigate a wide range of questions around the GR's impact on mortality and associated heterogeneity, robustness, and mechanisms.
2. Medicare (Section 3): This project also investigates some questions that can only be answered with Medicare data (for instance, around recent nursing home admission). Questions involving Medicare data are investigated using a separate, smaller pipeline.
3. Welfare (Section 4): Welfare analysis conducted in Section 5 of the paper is also run on a separate pipeline.
4. HRS data (Section 5): Lastly, we use restricted HRS data for a few small supplementary analyses, which is run under its own pipeline.

All of these pipelines share a basic common process: They take raw data, located in a separate location or locations, they clean this data (frequently in "build"), they analyze this data and output exhibits (frequently in "analysis"), and then these figures are added to the paper. Figures are mostly .pdf

files, while tables are most commonly exported in .txt files within this replication kit. The spreadsheet *exhibit_list.xlsx* provides the pipeline name, analysis folder name, and individual file name of every paper exhibit by exhibit number.

Some raw data is made available in this replication kit, while other raw data is restricted and cannot be shared; we articulate in the guide below which data can and cannot be publicly accessed.

Lastly, note that Section 6 is an appendix. Here, you can find a comprehensive glossary of acronyms used throughout this document.

2 NCHS pipeline

The large majority of analyses in this paper employ NCHS data, and are conducted within what we term the "NCHS pipeline." The structure and output of this data pipeline is displayed below. Note that some scripts use supplementary Stata functions found in the "ado_library" folder. *Note that while only the NCHS mortality data is restricted and thus excluded from the "raw_data" folder of this pipeline, this data is central to the pipeline, so most of the pipeline's outputs cannot be produced without first applying for and gaining access to NCHS mortality data. (And cleaned intermediate data is not provided due to cell-size restrictions by the CDC.)*

2.1 Data Sources: "raw_data" folder

2.1.1 Overview of most commonly used data by category

Most analyses are conducted using data that can be grouped into just three categories: mortality data, population data, and unemployment data. Most analyses include each of these in some form, with supplementary data sources sometimes added in to evaluate heterogeneity, mechanisms, and/or robustness.

- Mortality data: For mortality, we rely chiefly on data from the CDC, which is made available on a restricted basis by the National Center for Health Statistics (NCHS). This data contains individual-level records on deaths, including individuals' county of death, age, race, sex, and education level. (In analyses for the Medicare pipeline, we use Medicare data rather than CDC data for mortality).
- Population data: In order to generate mortality rates for a given geographic area (typcially a CZ or county), we also need to have access to that area's underlying population. We typically pull population data from the National Cancer Institute's Surveillance Epidemiology and End Results (SEER) program; SEER population estimates provide a slight modification of the US Census Bureau's intercensal population estimates.
- Unemployment data: We also need unemployment data to answer our question of interest. We can pull county-level unemployment rates and other employment data from the Bureau of Labor Statistics' Local Area Unemployment Statistics (LAUS); this is the same source previously used by Yagan (2019)[4] to estimate unemployment at the CZ-year level. In addition to using raw LAUS data for some lines of inquiry, we frequently pull unemployment estimates directly from the Yagan (2019)[4] replication package; these data are contained in raw data folders like *Yagan_replication_files* and *GR_county_shock*.

2.1.2 Inventory of raw data folders

Data folders are in alphabetical order within data category. Many data folders contain more detailed readmes with instructions for downloading the data they hold.

- Mortality / health data:

- *BRFSS*: The Behavioral Risk Factor Surveillance Survey (BRFSS) tracks morbidity and health behaviors for a surveyed population (accessible here). This data is used in the *Clean_BRFSS* build script, and ultimately the *BRFSS* analysis script, to produce a number of figures.
- *CDC_ageadj_mortality*: Contains the CDC's age adjusted death rates for 1900-2018 (accessible here), which are used to plot the long-run changes in mortality rates in the U.S. in the analysis script *historical_mortality*.
- *NCHS_RESTRICTED*: Our restricted-access individual-level mortality data comes from NCHS, and is used throughout our analyses; it is cleaned in build programs *clean_NCHS* and *collapse_NCHS*, and then used in most analysis programs. **It is restricted-access and not available for download; this is the only data source within the NCHS pipeline which cannot be downloaded**. Instructions for applying for this data can be found here, specifically, we use the data file "Deaths (Mortality) – Multiple cause of death, states and all counties – Detailed" (description here).
- *NCHS_Unrestricted*: Individual-level mortality data from 1968-1988, used to construct age-adjusted mortality rates in *collapse_NCHS_early* and then aggregated to a standardized national mortality rate in *NCHS_descriptive* and used to plot the long-run mortality time trend against unemployment. Data is downloaded from (here).

- Population /demographic data:
  - *ACS_new*: The American Community Survey (ACS) is an annual survey run by the Census, which collects information including individuals' age, education, and location. This data is therefore useful in assessing mortality rates within education groups, as our standard population denominator, SEER, does not collect education information. This data is therefore used in *NCHS_by_educ* (and *demographic_balance* and *population_composition*) to compute education-specific mortality estimates. For some additional analyses in the analysis script *recovery_compare*, we again use the ACS survey for a set of labor force participation variables. This data can be pulled here.
  - *Census_county_population*: Generally, we use SEER data for our population denominator, at least as relates to mortality. However, when Yagan (2019)[4] calculates unemployment, they combine LAUS employment data with Census population data for their denominator. When we construct additional economic variables (like employment-to-population ratio, or EPOP), we maintain this same procedure, merging Census population data with LAUS data in the build script *LAUS_data*, which is subsequently used in the analysis programs *EPOP_recovery*, *NCHS_robust_fracking*, and *recovery_compare*, each of which examines EPOP. Data can be pulled here for 2000-2009 and here for 2010-2019.
  - *NHIS*: Contains individuals' year of death, gender, race, marital status, and education 2003-2016, used in the analysis script *demographic_balance* to predict an individual's probability of death in a given year based on those demographic characteristics. Data extract comes from IPUMS and can be pulled here.
  - *urban_rural*: For some additional analyses in *recovery_compare*, we again use the ACS survey, specifically pulling an urban/rural indicator variable that had not been previously used. This data is stored separately, but can also be pulled here.
  - *US_county_population_seer*: Contains data from the SEER with yearly county-level data on population, including by race, gender, and age. This data is cleaned in *clean_SEER* and then used repeatedly throughout analysis. Data is retrieved from the SEER website here.

- Unemployment /economic data:
  - *ADH_manufacturing*: In the analysis script *recovery_compare*, we also pull data on CZ-level manufacturing share of employment from the Autor, Dorn, and Hanson (2016)[1] replication kit (accessible here). This data was originally derived from the Census and the ACS.
  - *CPI*: When evaluating the GR's impact on earnings in *CPS_analysis*, we normalize to 2015 dollars. Doing so requires CPI data, which we can pull from the Minneapolis Fed here.

- *CPS*: As mentioned directly above, the analysis in *CPS_analysis* uses data from CPS, or the Current Population Survey, to gather information on earnings. We can pull this data from IPUMS here.
- *CPS_unemp_natl*: Contains data on national unemployment from the 1950s to present. Used to plot long-run unemployment time trend against mortality rate in *NCHS_descriptive*. Data is downloaded from here.
- *GR_county_shock*: Contains data with 2007-2009 Great Recession unemployment shocks (GR shocks) at the commuting zone level, used in many build and analysis scripts. (Note that it misleadingly has "county" in its name, but is actually used for CZ-level shocks. More confusingly yet, a script in build called *county_GR_shock* generates shocks that are actually at the county level). Data is retrieved from Yagan (2019)[4] replication package, and can be downloaded here.
- *housing_prices*: Contains county and national level House Price Index data obtained from the Federal Housing Finance Agency (accessible here), and serves as an input to the analysis script *recession_measures*.
- *LAUS_data*: As is described above when discussing Census_county_population, Yagan (2019)[4] uses LAUS data to generate unemployment shocks, and we typically use Yagan's constructions directly (as can be found, for instance, in *GR_county_shock*). However, we occasionally use the raw LAUS data directly, particularly in analyses around EPOP. This data is retrieved from here.
- *US_county_employment*: Contains counties' year-level unemployment and labor force statistics between 1999-2020, drawn from LAUS. This data is used in the build program *county_GR_shock* to develop 2007-2009 unemployment shocks on the county level, which is subsequently used in analysis programs *NCHS_robust_geo* and *pm2dot5_analysis*. Data is retrieved from the Bureau of Labor Statistics here.
- *US_county_output_BEA*: Contains 2001-2021 county-year real GDP from the Bureau of Economic Analysis (CAGDP1 time series, accessible here) that feeds into the build script *clean_output_BEA*, which then feeds into the analysis script *recession_measures*.
- *Yagan_replication_files*: Beyond the Yagan (2019)[4] replication data on CZ-level unemployment which is available in *GR_county_shock*, we occasionally use a range of other replication data, particularly in the analysis scripts *NCHS_by_educ* and *NCHS_robust_geo*. This data is retrieved from here.

- Pollution data:
  - *CensusBlockGroupShapes*: Contains 2010 Census Block Group (CBG) shapefiles, used to map CBGs to grid squares of pollution and wind direction measurements in the build scripts *satellite_pm2dot5* and *prevailings_winds*. Data is downloaded separately for each state from here.
  - *Census_CBG_population*: Contains 2010 Census Block Group (CBG) populations, which are used to weight pollution and wind direction measured over grid squares by the populations of those grid squares in the build scripts *satellite_pm2dot5* and *prevailings_winds*. Data is downloaded from here.
  - *Pollution_AQS*: This folder holds the EPA's Air Quality System (AQS) data, which serves as a secondary source of geographic pollution data. Though this data is cleaned in the build script *pollution_cleaning,* the output of which is sparsely used in some secondary analyses, this data is mainly used directly as an input to *pm2dot5_analysis*, where the majority of pollution-related data cleaning and analysis is conducted. This data is retrieved from here.
  - *Pollution_Satellite*: This is the source for the PM2.5 concentration estimates covering the entire U.S. from 2003-2016: van Donkelaar et al. (2020) produces estimates based on satellite data, corrects them based on a chemical transport model (structural model that uses the laws of physics as its assumptions), and validates them against ground monitor data. The data is retrieved from here. This data is cleaned in *satellite_pm2dot5* and then used as an input to *pm2dot5_analysis* and *wind_direction_iv*.

  - *wind_direction*: This is the source of the monthly averaged wind direction data from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 from 2003-2016. The data is retrieved from here. This data is cleaned in *prevailing_winds* and then used as an input to *wind_direction_iv*.

- Other data:
  - *crosswalks*: Contains a range of crosswalks between geographies which we use repeatedly throughout analyses. This data is retrieved from here.
  - *fracking*: The analysis script *NCHS_robust_fracking* looks specifically at the effect of the GR on unemployment when removing high-fracking geographies. Data on high-fracking counties is taken from the replication package of Bartik et al. (2019)[2], and can be downloaded here.
  - *lifetables*: Contains actuarial life tables, separately for men and women, for each year between 1900 and 2019. This data is used in the analysis script *life_tables* to estimate the impact of the GR on life expectancy. Data is retrieved from the Social Security Administration (SSA), and can be downloaded here.

2.2 Data cleaning: "build" folder

In the "build" folder, we clean the data that will ultimately be used to conduct analysis. In Figure 1 (Section 2.2.2), you can see a flowchart of these build scripts in yellow, with blue circles indicating raw data files and arrows indicating a build file's outputted clean data being used in a subsequent script. Table 1 (Section 2.2.2) also shows which build outputs are used in each analysis script (all are used in at least one analysis script except outputs of *clean_NCHS_early* and *clean_SEER_early*, which are only used for subsequent build scripts). Every script in the "build" folder outputs its cleaned data to a folder sharing the same name within the "clean_data" folder.

2.2.1 Written descriptions of scripts

Below is a brief description of the pipeline's cleaning scripts:

- *clean_BRFSS*: Cleans raw BRFSS data, which is subsequently used in the analysis script *BRFSS* to examine the impact of the Great Recession on morbidity and health behaviors.
- *clean_output_BEA*: Cleans raw BEA local GDP data, which is subsequently used in the analysis script *recession_measures*.
- *clean_NCHS*: Cleans raw (restricted) NCHS data on mortality.
- *clean_NCHS*_early: Cleans raw (unrestricted) NCHS data on mortality for earlier years of analysis.
- *clean_SEER*: Cleans raw SEER data on geography-year population counts. Feeds into *collapse_NCHS* and is also used directly in a range of analysis scripts.
- *clean_SEER*_early: Cleans raw SEER data on geography-year population counts for earlier years of analysis.
- *collapse_NCHS*: Combines output from *clean_NCHS* (mortality) and *clean_SEER* (population) to produce estimates of mortality rates (including within demographic groups and among particular causes of death). Is later used throughout a wide range of analysis scripts.
- *collapse_NCHS_early*: Fulfills a similar function to *collapse_NCHS* for earlier years of data.
- *collapse_NCHS_deaths_of_despair*: Performs a similar function to collapse_NCHS, combining data from *clean_NCHS* and *clean_SEER*. However, this script estimates mortality rates specifically for deaths of despair, which will later be used in analysis script *NCHS_deaths_of_despair*.

- *county_GR_shock*: Uses county-level employment data to produce estimates of the 2007-2009 unemployment shock at the county level, used across a range of sensitivity analyses that look at the county level. (Note that the raw data in *GR_county_shock* measures the 2007-2009 unemployment shock at the *commuting zone* level and serves as the paper's primary shock measure.)

- *LAUS_data*: Uses county-level population estimates and LAUS labor force statistics to produce geographic estimates of labor force participation (particularly EPOP) using a similar procedure to Yagan (2019)[4].

- *pollution_cleaning*: Uses pollution data from *Pollution_AQS*, as well as CZ-level shocks (*Yagan_replication_files*), county-level shocks (*county_GR_shock*), population data (*clean_SEER*) and mortality data (*collapse_NCHS*) to clean AQS-based pollution data for analysis.

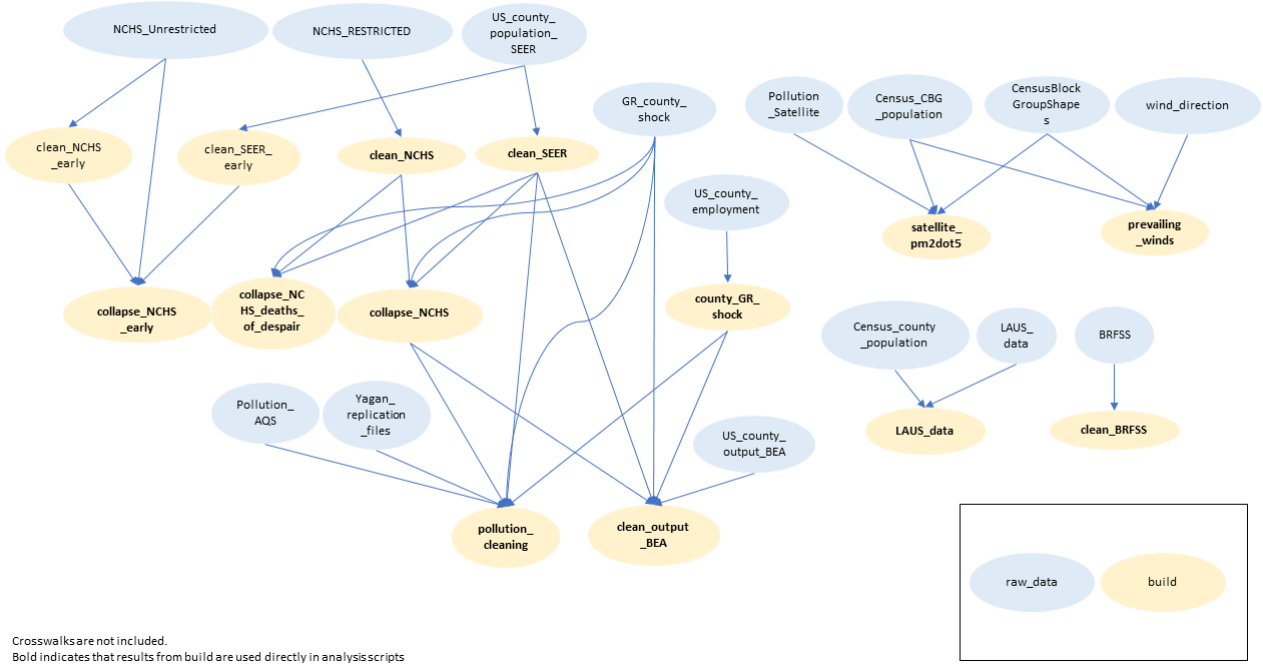
- *satellite_pm2dot5*: Cleans satellite-based pollution data from *Pollution_Satellite*, weighting using regional population data from *Census_CBG_population* and shapefiles from *CensusBlockGroupShapes*. (There is a .R script with this name that extracts data from the raw file formats and needs to be run before the .do file with this name.)

- *prevailing_winds*: Cleans data on prevailing wind direction from *wind_direction*, weighting using regional population data from *Census_CBG_population* and shapefiles from *CensusBlockGroupShapes*. (There is a .R script with this name that extracts data from the raw file formats and needs to be run before the .do file with this name.)

2.2.2 Relevant figures and tables

This section shares two exhibits: Figure 1 links raw data files to build scripts; Table 1 and Table 2 link clean data folders (which share names with the build scripts that produce the clean data) to analysis scripts (split into two tables just for sizing/formatting reasons).

Figure 1: Linkages between raw data and build folders



Note: This flowchart shows the pipeline for build scripts. It can be interpreted sequentially from top to bottom. All build scripts are colored in yellow. Raw data is colored in blue. Crosswalk data is not included to reduce clutter, but comes into frequent use. Build scripts whose output (which is located in a folder sharing the script name within the "clean_data" folder) is pulled directly into analysis folders are bolded.

Table 1: Linkages between build scripts and analysis scripts (Pt 1)

Analysis folder	**Build folder**								
	Collapse _NCHS	Clean _NCHS	Collapse _NCHS_ deaths_of _despair	Collapse _NCHS _early	Clean _SEER	LAUS _data	Clean_ BRFSS	Clean _output _BEA	county_ GR_shock
alt_cause_of_death	X	X			X				
BRFSS					X		X		
CPS_analysis					X				
demographic_balance	X								
EPOP_recovery	X					X			
*fill_tables									
historical_mortality									
life_tables									
linearity_check	X								
NCHS_by_educ	X	X			X				
NCHS_deaths_of_despair			X		X				
NCHS_decompositions	X				X				
*NCHS_descriptive	X			X	X				
*NCHS_mortality	X								
NCHS_robust_fracking	X					X			
NCHS_robust_geo	X								X
NCHS_robust_outliers	X								
NCHS_robust_percent	X					X			
NCHS_robust_quartiles	X								
NCHS_robust_sunshapiro	X								
population_composition					X				
recession_measures					X	X		X	
recovery_compare	X					X			

Note: Each row corresponds to an analysis script, and each column corresponds to a build script. An X is marked in a cell if a given analysis script uses cleaned data output from a given build script (which is located in a folder sharing the script name within the "clean_data" folder). (A flowchart would be too cluttered to illustrating all the connections.) * Indicates that an analysis script relies on output produced by another analysis script.

Table 2: Linkages between build scripts and analysis scripts (Pt 2)

Analysis folder	**Build folder**					
	Collapse _NCHS	Clean _SEER	pollution _cleaning	satellite _pm2dot5	prevailing _winds	county_ GR_shock
pm2dot5_analysis	X	X	X	X		X
wind_direction_iv	X	X		X	X	X

Note: Each row corresponds to an analysis script, and each column corresponds to a build script. An X is marked in a cell if a given analysis script uses cleaned data output from a given build script (which is located in a folder sharing the script name within the "clean_data" folder). (Pollution analysis is its own table for table sizing purposes, as several of the pollution data sources are only used in the two above analysis scripts.)

2.3 "Analysis" folders

Note on conventions for referencing analysis scripts in this readme: Because analysis on particular topics is sometimes spread across multiple code files for clarity (ex. *wind_direction_iv_build,* *wind_direction_iv_first_stage,* and *wind_direction_iv_mediation*), all three scripts are grouped under the analysis script group (the first part of the file name) *wind_direction_iv* and are referred to as

wind_direction_iv throughout. However, most analysis script groups only have one associated script, and end with the suffix _analysis. For example, the script named *NCHS_by_educ_analysis* is referred to as *NCHS_by_educ*, and no other analysis scripts' names begin with *NCHS_by_educ. The order that scripts within an analysis script group should be run is specified in the* make.py *file that runs all code within the analysis/code folder.*

The "analysis" scripts produce the paper exhibits. Below is a brief description of each analysis script group, as well as the figures in the draft that they produce.

These script groups are organized relatively horizontally: With relatively few exceptions, they can all be run simultaneously, once build scripts are run. The three exceptions where an analysis script group pulls directly from another analysis script group (*fill_tables*, *NCHS_descriptive*, and *NCHS_mortality*) are laid out clearly below.

A flowchart of the raw and cleaned data that is used for each analysis script group would be too visually convoluted to be helpful. But links between build (output located in "clean_data") and analysis script groups can be found in Figure 1 (and running "command F" for "/raw_data/" in an analysis script is a quick way to see the raw data inputs into a particular analysis script).

The analysis script groups are organized into several broad categories, in terms of their functions:

- Mortality: produces the main event study that evaluates the impact of the Great Recession shock on mortality, particularly among race, age, education, and gender subgroups, as well as among different causes of death
- Sensitivity: limits our results to certain geographies, pursues different identification strategies, defines unemployment shock differently, assesses linearity
- Mechanisms: runs supplementary event studies where the outcome is pollution, health behaviors, or earnings
- Aggregation: describes broad trends in unemployment, mortality, pollution, or other variables of interest

2.3.1 Main mortality event studies

- *NCHS_by_educ*: Examines heterogeneity in main event study findings by education level, using ACS data to construct the population denominator (as SEER data does not contain population by education level). Creates Figure OA.18, Figure OA.42, and Figure OA.43.
- *NCHS_decompositions*: Provides the share of mortality reductions concentrated in various age group and cause of death categories. Creates Figure IV (b), Figure VI (b), and Figure OA.52.
- *NCHS_mortality*: Produces main event study for the impact of the GR on mortality, heterogeneity by gender, race, age, and cause of death, and some variants on functional form (including non-logged mortality and poisson). Also produces plots comparing period estimates for the GR impact across a range of subgroups. *NCHS_mortality_analysis.do* produces the main event studies, while *NCHS_mortality_period_estimate_exhibits.do* produces coefficient plots. Script pulls in *NCHS_by_educ* for coefficient plots, meaning this script group needs to be run after *NCHS_by_educ*. Creates Figure III, Figure IV (a), Figure V, Figure VI (a), Figure VII, Figure OA.7, Figure OA.8, Figure OA.17 (a), Figure OA.22 (a), Figure OA.27 (a) and (b), Figure OA.40, Figure OA.41, Figure OA.45, and Figure OA.46.

2.3.2 Sensitivity analyses

- *alt_cause_of_death*: Examines robustness of findings on the heterogeneity of the GR effect by cause of death for alternative cause-of-death categorizations. Creates Figure OA.9, Table OA.12, Table OA.13, and Table OA.14.

- *demographic_balance*: Examines robustness of main findings to secular trends in mortality by gender, race, marital status, and education. Creates Figure OA.14.
- *epop_recovery*: Provides summary statistics and estimates the impact of the GR using shocks in EPOP rather than the unemployment rate. EPOP data is drawn from LAUS. Creates Figure OA.22 (b), Figure OA.23, and Figure OA.25.
- *linearity_check*: Plots Great Recession shock against change in mortality as a test of linearity, using bin scatter. Creates Figure OA.28.
- *NCHS_robust_fracking*: Using data from Bartik et al (2019)[2], examines robustness of findings to the removal of high-fracking CZs. Creates Figure OA.26 (e).
- *NCHS_robust_geo*: Checks robustness of main findings to using different geographic units (like counties or states) rather than CZs. *NCHS_robust_geo_county_analysis.do* looks at the county level, *NCHS_robust_geo_state_analysis.do* looks at the state level, and *NCHS_robust_geo_state_by_year.do* adds Census-division-by-year controls. Creates Figure OA.26 (a) and (b) and Figure OA.27 (c).
- *NCHS_robust_outliers*: Checks robustness of main findings to removing potentially influential geographic outliers. Creates Figure OA.26 (c) and (d).
- *NCHS_robust_percent*: Checks robustness of main findings to defining the shock in percent change rather than percentage point change. Creates Figure OA.29.
- *NCHS_robust_quartiles*: Checks robustness of main findings to dividing CZs into quartiles of unemployment shock, rather than treating unemployment shock as a continuous variable. Creates Figure OA.27 (d), (e), and (f).
- *NCHS_robust_sunshapiro*: Inspired by Sun and Shapiro (2022)[3], performs a series of imputations meant to evaluate whether treatment effect heterogeneity may be biasing results (for instance, if CZs with larger unemployment shocks also tend to have different marginal effects of unemployment shock on mortality). Creates Figure OA.30.

2.3.3 Mechanisms:

- *BRFSS*: Uses data from BRFSS to examine the impact of the Great Recession on morbidity and health behaviors. Creates Figure VIII, Figure OA.48, Figure OA.49, Figure OA.50, Figure OA.51, Table OA.4, and Table OA.10.
- *CPS_analysis*: Uses data from CPS to estimate the impact of the Great Recession on earnings, using state-level OLS and individual-level poisson, and estimating separately among those in different education and nativity age brackets. Creates Figure OA.20.
- *NCHS_deaths_of_despair*: Provides a more in-depth look at trends around deaths of despair. Creates Figure OA.11.
- *pm2dot5_analysis*: Performs all analyses on pollution data, except work related to the wind instrument—which is done in *wind_direction_iv*. Pollution data is first cleaned in *pm2dot5_analysis_build.do*; descriptive statistics are generated in *pm2dot5_analysis_descriptive_stats.do*; and finally, the main event studies are run in *pm2dot5_analysis_event_studies.do*. Creates Figure IX, Figure OA.32, Figure OA.33, Figure OA.34, and Figure OA.35.
- *wind_direction_iv*: Examines prevailing wind direction as an instrument for pollution for some supplementary analyses. The instrument is constructed in *wind_direction_iv_build.do*, the first stage is evaluated in *wind_direction_iv_first_stage.do*, and mediation analyses are run in *wind_direction_iv_mediation.do*. Creates Figure OA.36, Figure OA.38, and Table OA.19. (Note that Figure OA.37 is also associated with this analysis. However, this figure was created in Photoshop for illustration purposes only and contains no data or generated estimates, so this figure is beyond the scope of this replication package—the only paper exhibit where this is the case.)

2.3.4 Aggregation / descriptive:

- *fill_tables*: This script pulls together output from a range of other analysis scripts and consolidates them. Therefore, this script should be run last. Creates Table OA.15, and Table OA.18.
- *historical_mortality*: Plots mortality over time (from 1956 to 2006), drawn entirely from CDC death rate data. Creates Figure OA.10.
- *life_tables*: Using actuarial life tables from the SSA, produces tables which estimate the GR's impact on life expectancy. Creates Table OA.2.
- *NCHS_descriptive*: Provides a range of figures that lay out summary statistics around mortality (e.g. scatterplots, geographic heat maps). Separate .do files create main analyses (*NCHS_descriptive_analysis.do*), produce maps (*NCHS_descriptive_maps*.do), and check for instances of zero mortality (*NCHS_descriptive_check_zeros.do*). Also produces some tables, including by pulling output from *NCHS_by_educ* and *NCHS_mortality*, meaning that it has to be run after those scripts. Creates Figure I, Figure II, Figure OA.1, Figure OA.5, Table OA.1, and Table OA.3
- *population_composition*: Produces event studies on the impact of the GR on population quantity and composition. Creates Figure OA.12 and Figure OA.13.
- *recession_measures*: Produces analyses around various potential measures of recession severity, particularly EPOP, GDP per capita, and HPI. *recession_measures_cz_analysis.do* estimates event studies while *recession_measures_descriptive_stats.do* generates summary statistics. Currently corresponds to Figure OA.2, Figure OA.3, and Figure OA.4.
- *recovery_compare*: Produces figures evaluating differences between communities with high vs low recession recoveries. Creates Figure OA.24 and Table OA.11.

2.4 "Output" folders

The names of the output folders correspond exactly with the names of the analysis script group folders described above in Section 2.3. All figures appear within a "figures" sub-folder, and all tables appear within a "tables" sub-folder. Note that there may be a few extraneous figures in some "figures" folders that are not included in the final paper draft. (*exhibit_list.xlsx* links each exhibit from its exhibit number to the output folder where it is located and the file name it is called.) The "matrices" sub-folders contain coefficient estimates, some of which get compiled together into tables, but no paper exhibits are located in a "matrices" folder.

3 Medicare pipeline

A portion of our data pipeline is run separately from the main pipeline as described in Section 2. The pipeline detailed below relates overwhelmingly to the cleaning and analysis of Medicare data, and was originally constructed within the NBER's aging server cluster. Within this portion of the project, Medicare data is both used for our death numerator (substituting for NCHS data), and often for our population denominator. Due to data access restrictions, the work done within this portion of the pipeline is entirely self-contained.

3.1 Data Sources: "raw_data" and "restricted_data" folders

3.1.1 Overview of most commonly used data

Most of the data used here is therefore Medicare data, including data on claims, chronic conditions, recipient demographics, and inpatient and SNF stays. All such data is kept in the "restricted_data" folder; and cannot be provided publicly.

This data is supplemented with a range of public-access and researcher-collected data, which is kept in the "raw_data" folder and can be shared publicly. Much of this data is duplicated from that described in Section 2.1.

3.1.2 Inventory of raw data folders

"restricted_data" folder:

The "restricted_data" folder contains a range of raw Medicare data, typically adopting the 20% representative sample to reduce runtime. This data is restricted-access and cannot be shared publicly. See more information on these data sources and how to request access within the relevant readme files.

- *bsfcc20*: This folder contains Chronic Conditions data (20% sample), which reports whether Medicare beneficiaries have one of 27 chronic conditions in a given year. This data is cleaned in *Chronic_Conditions*, which then is used to estimate counterfactual life expectancy for use in *patient_summary_statistics* and *life_years_analysis*.
- *bsfcc20_h*: This folder contains Chronic Conditions data (100% sample), which is used for years in which the 20% sample is unavailable. The data is ultimately used in the same context as for *bsfcc20*, and is similarly cleaned in *Chronic_Conditions*.
- *claims*: This folder contains Medicare Centralized Claims data (20% sample) for Medicare Parts A and B, including inpatient, outpatient, and carrier files. After being cleaned in *Utilization_Indicators*, we use this data to construct a dataset with utilization information in *Collapsed_Data_Utilization*, which is ultimately used in *utilization_analysis_cz* to evaluate the impact of the GR on measures of utilization like total expenditure, physician visits, ER visits, and inpatient visits.
- *denom20*: These Medicare Centralized Denominator files (20% sample) contain recipient-level information on Medicare beneficiaries' demographics (including race, age, sex, and Medicaid status). We clean this data in *Baseline Sample* and *Individual Data Cleaning*, which then ultimately filters into all of our analysis work.
- *ehic_bene_xwalk*: This folder contains a crosswalk between antiquated ("ehic") and current ("bene_id") identifiers for Medicare beneficiaries, used in *SNF_Utilization* as part of the data cleaning process.
- *medpar*: MedPAR data (100% sample) contains information about inpatient hospital and SNF stays that were covered by Medicare. We clean this data in *SNF_Utilization* and *Collapsed_Data_by_SNF*, and ultimately use the data in *shock_mortality* to look at heterogeneity by recent SNF stay.

"raw_data" folder:

The "raw_data" folder contains a range of supplementary, non-Medicare data which is either publicly available or was generated directly by researchers. It is included directly as part of the replication kit. See more information on how to download these data from the relevant readme files.

- *cz_crosswalk*: Contains crosswalks from CZs to counties and Census divisions, used in several build and analysis scripts.
- *economist_survey*: Contains survey data collected by researchers regarding the impact of the Great Recession on mortality. Cleaned and analyzed in *survey*.
- *geo*: Contains a range of additional geographic information; used during data cleaning for *Baseline_Sample*.
- *GR_shock*: Contains data from the Yagan (2019)[4] replication package with unemployment shocks at the CZ and state level, needed for the bulk of analyses.

- *OSCAR_Facilities*: We use data from the Online Survey Certification and Reporting (OSCAR) database, which comprise facility-level inspection-based administrative data, to measure nursing home staffing. This raw data is fed directly into the analysis script *shock_OSCAR*, which produces figures on nursing home staffing and occupancy characteristics.

- *seer_raw*: SEER microdata shares yearly county-level population data, including by race, gender, and age. This is the same underlying data that is used for population-related analyses for the NCHS pipeline. It is cleaned in *SEER_Data* and then used in several analysis scripts.

3.2 Data cleaning: "build" folder

In the "build" folder, we clean the data that will ultimately be used to conduct analysis. In Figure 2 (Section 3.2.2), you can see a flowchart of these build scripts in yellow, with blue circles indicating raw data files, green circles indicating restricted raw data files, and arrows indicating data being used in a subsequent script. Every script in the "build" folder outputs its cleaned data to a folder sharing the same name within the "cleaned_data" folder.

Note that some scripts use supplementary Stata functions found in the "ado_library" folder.

3.2.1 Written descriptions of scripts

Below is a brief description of the pipeline's cleaning scripts:

- *Baseline_Sample*: Cleans Denominator data for use in *Individual_Data_Cleaning*. Is also used directly in *fill_tables*.

- *Chronic_Conditions*: Cleans Chronic Conditions data, primarily for use computing counterfactual life expectancy. Is then used in *Collapsed_Data_Utilization* and *Life_Years_Lost*.

- *Collapsed_Data_by_SNF*: Taking cleaned Denominator data from *Individual_Data_Cleaning* and SNF data from *SNF_Utilization*, this program produces individual-year data on qualities of SNF care received, as well as CZ-year-level data on mortality among SNF utilizers and non-SNF-utilizers. Ultimately used in *shock_mortality* and *numbers_in_text*.

- *Collapsed_Data_CZ*: Taking cleaned Denominator data from *Individual_Data_Cleaning*, produces CZ-year-level data on mortality among several subsamples of Medicare enrollees, used in *shock_mortality, numbers_in_text,* and *analysis_by_xs*.

- *Collapsed_Data_Utilization*: Pulling cleaned Claims and Chronic Conditions data from *Utilization_Indicators* and *Chronic_Conditions*, respectively, creates individual-year level dataset on individuals' chronic conditions and receival of care, ultimately used in *utilization_analysis_cz*.

- *Individual_Data_Cleaning*: Receives data from *Baseline_Sample*, and further cleans Denominator data for use across a broad range of build and analysis programs.

- *Life_Years_Lost*: Pulling cleaned Denominator data from *Individual_Data_Cleaning* and Chronic Conditions data from *Chronic_Conditions*, uses a Gompertz specification to predict life expectancy for all patient-years (using separate scripts in Stata and R, which are then employed in *patient_summary_statistics* and *life_years_analysis*.)

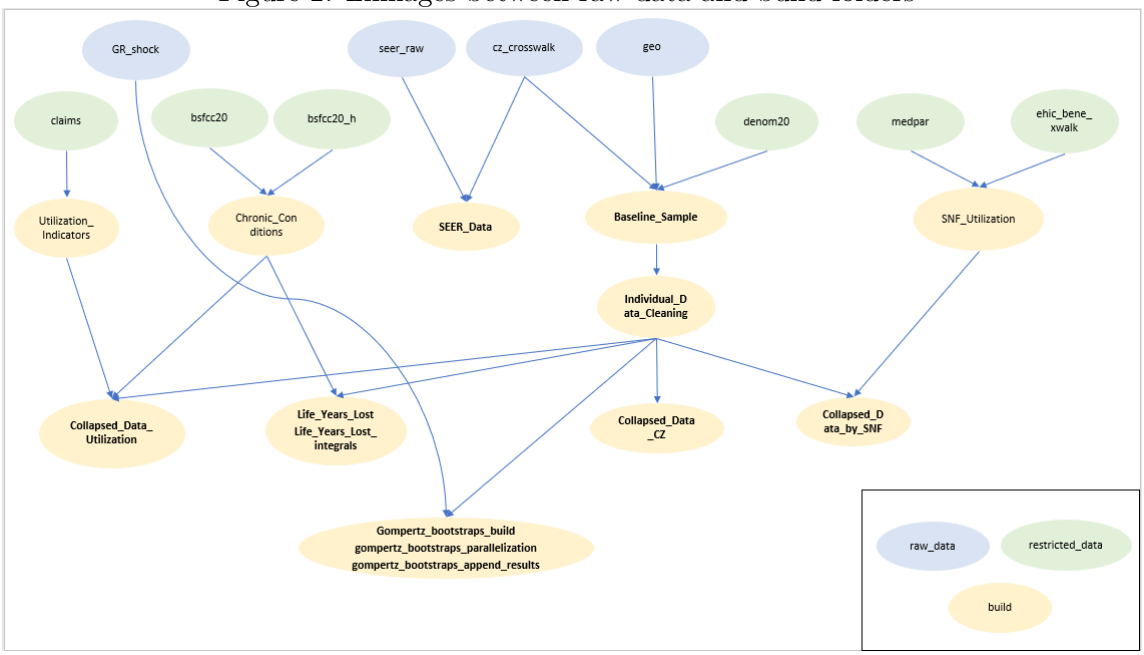
- *gompertz_bootstraps*: In order to run *gompertz_analysis*, we need to calculate standard errors for our control function. This is done with a Bayesian bootstrap procedure conducted in this script, which contains several distinct Stata programs with a cumulative runtime of at least a week. Results are used for *gompertz_analysis*; the data needed to run the script are drawn from *Individual_Data_Cleaning*.

- *SEER_Data*: Cleans SEER-based population data for use in a broad range of subsequent analysis scripts.

- *SNF_Utilization*: Cleans MedPAR data to produce individual-year level data on total SNF expenditures and total number of days within a SNF, to be used in *Collapsed_Data_by_SNF* and ultimately *shock_mortality*.
- *Utilization_Indicators*: Cleans Claims data for later use in *Collapsed_Data_Utilization*.

3.2.2 Relevant figures and tables

Figure 2: Linkages between raw data and build folders



Note: This flowchart shows the pipeline for build scripts. It can be interpreted sequentially from top to bottom. All build scripts are colored in yellow. Raw data is colored in blue. Restricted raw data is colored in green. Build scripts whose output (which is located in a file sharing the script name within the "cleaned_data" folder) is pulled directly into analysis folders are bolded.

3.3 Data analysis: "analysis" folder

The "analysis" scripts produce the paper exhibits. Below is a brief description of each analysis script group, as well as the figures in the draft that they produce.

These script groups are organized relatively horizontally: With relatively few exceptions, they can all be run simultaneously, once build scripts are run. The two exceptions where an analysis script group pulls directly from another analysis script group (*fill_tables* and *snf_comparefig*) are laid out clearly below.

In Figure 3 (Section 3.3.5), you can see a flowchart of these analysis scripts in gray, with blue circles indicating raw data files, yellow circles indicating cleaned data pulled from build scripts, orange arrows indicating final paper outputs, and arrows indicating data being used in a subsequent script. Every script in the "analysis" folder outputs its exhibits to a folder sharing the same name within the "output" folder.

The analysis script groups are organized into several broad categories, in terms of their functions:

- Mortality: produces the main event study that evaluates the impact of the Great Recession shock on mortality, particularly among race, age, education, and gender subgroups, as well as among different causes of death
- Sensitivity: limits our results to certain geographies, pursues different identification strategies, defines unemployment shock differently, assesses linearity

- Mechanisms: runs supplementary event studies where the outcome is pollution, health behaviors, or earnings
- Aggregation: describes broad trends in unemployment, mortality, pollution, or other variables of interest

Note that some scripts use supplementary Stata functions found in the "ado_library" folder.

3.3.1 Mortality event studies, overall and by subgroup:

- *analysis_by_xs*: This script runs our main event study for subsamples of our Medicare population who are and who are not on Medicaid. Creates Figure OA.44.
- *shock_mortality*: This script runs our main event study for a range of different subsamples, including individuals receiving and not receiving SNF care, and several definitions of our Medicare population (overall, receiving TM in t, and receiving TM in t-1). Creates Figure OA.17 (b), (c), and (d), Figure OA.19, and Figure OA.53. Results are also used by *fill_tables* to produce Column (1) of Tables OA.16 and OA.17.

3.3.2 Sensitivity analyses:

- *gompertz_analysis*: We use a Gompertz specification to estimate the impact of population flows on our main findings, looking at the Medicare population. Creates Figures OA.15 and OA.16. Results are also used by *fill_tables* to produce Table I.
- *life_years_analysis*: Conducts analysis on the life years lost that were prevented by the GR, under a range of assumptions around counterfactual life expectancy. This was undertaken to evaluate the possibility of the marginal life saved by the GR having a very low counterfactual life expectancy. Does not directly produce any figures in the draft, but its results are used by *fill_tables* to produce Columns (2)-(5) of Tables OA.16 and OA.17.

3.3.3 Mechanisms:

- *shock_OSCAR*: Uses data from the OSCAR survey to evaluate the effect of the GR on characteristics of nursing home care. Creates Figures OA.54 and OA.55.
- *utilization_analysis_cz*: Evaluates the effect of the GR on measures of healthcare utilization, as contained within Medicare data. Creates Figure OA.21.

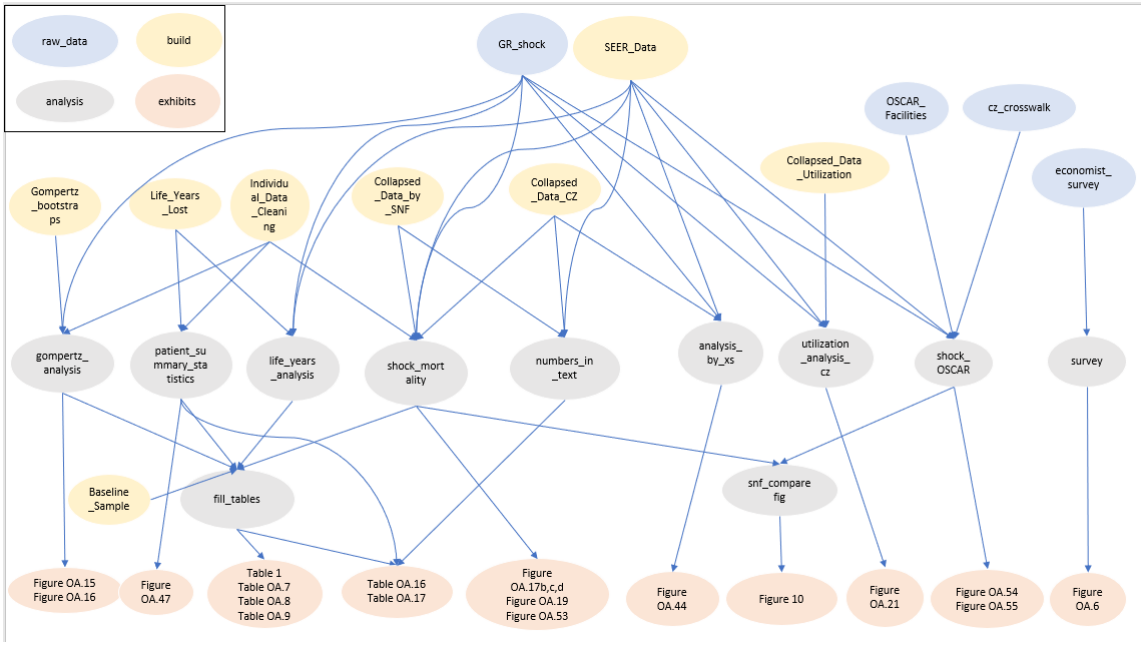
3.3.4 Aggregation / descriptive:

- *fill_tables*: Gathers results from a range of other analysis scripts (namely, *gompertz_analysis*, *shock_mortality*, *life_years_analysis*, and *patient_summary_statistics*, as well as the build script *Baseline_Sample*), and organizes them into clean, formatted tables for the draft. Produces Table I, Table OA.7, Table OA.8, Table OA.9, Table OA.16, and Table OA.17.
- *numbers_in_text*: This script produces a *number_in_text.tex* file, which contains the "Mean Mortality Rate (per 100,000)" statistic used in Tables OA.17 and OA.18
- *patient_summary_statistics*: Produces summary statistics on patients' demographics and their predicted average life expectancy, using data from the build script *life_years_lost* for the latter. Produces Figure OA.47, and its results are used by *fill_tables* to produce Table OA.9. Additionally, this script produces a *number_in_text.tex* file, which contains the "Mean LYL per Decedent" statistics used in Tables OA.17 and OA.18

- *snf_comparefig*: Uses output from the analysis scripts *shock_mortality* and *shock_oscar* to produce a figure laying out event study results related to SNF care. Produces Figure X.
- *survey*: Analyzes survey data from economists who were asked to predict the impact of the GR on mortality. Pulls directly from survey data in *economist_survey*. Produces Figure OA.6.

3.3.5 Relevant figures and tables

Figure 3: Linkages between raw data, build folders, and analysis folders



Note: This flowchart shows the pipeline for analysis scripts. It can be interpreted sequentially from top to bottom. All analysis scripts are colored in gray. Build scripts are colored in yellow, raw data is colored in blue, and final exhibits are colored in orange.

3.4 "Output" folder

The names of the output folders correspond exactly with the names of the analysis folders described in Section 3.3. All figures appear within a "figures" sub-folder, and all tables appear within a "matrices" sub-folder. (*exhibit_list.xlsx* links each exhibit from its exhibit number to the output folder where it is located and the file name it is called.)

4 Welfare pipeline

There is a range of macroeconomic analysis around the projected consumption impacts of recessions in general and the Great Recession in particular. These analyses are conducted within their own data pipeline.

4.1 Data

Raw and cleaned data are both stored within the "data" sub-folder. A readme within the folder gives instructions for downloading the raw data and links cleaned data to source raw data and cleaning scripts

4.1.1 Raw Data

- IPUMS Displaced Worker Supplement extracts (downloadable here)

  - *dws_tenure.csv*
  - *ipums_cps_dws_by_educ.csv*

- Mortality rates (separate for males and females, downloadable here)
  - *PerLifeTables_F_Hist_TR2022.csv*
  - *PerLifeTables_M_Hist_TR2022.csv*

- Population (downloadable here)
  - *us.1969_2020.singleages.adjusted.txt*

4.1.2 Cleaned Data

- *mortality_rates_educ.xlsx*
  - Sourced from mortality rate raw data, cleaned by script *mortality_rates_educ.do*
- *seer_population.csv*
  - Sourced from population raw data, cleaned by script *import_seer_population.R*

4.2 Code

4.2.1 Reports

Reports are longer code files, and they produce exhibit outputs.

- *extension_mortalities.Rmd*: baseline estimate of welfare cost of recessions accounting for versus not accounting for mortality, creates Figure XI (a) and Table OA.5
- *va-197.Rmd*: extends baseline results examining heterogeneity by mortality, *must be run after* extension_mortalities.Rmd, creates Figure XI (b)
- *extension_gr_pl.Rmd*: extends baseline results by accounting for retirement, creates Figure XI (c) and Table OA.6
- *va-98.Rmd*: extends baseline results by allowing for mortality effects of job displacements found in prior literature, creates Figure OA.39
- *vsly_parameters_table.R*: generates parameters used in welfare model calibrations, creates Table OA.20

4.2.2 Scripts

Scripts are smaller code files, the building blocks called by the report code files to perform various tasks. The two data cleaning scripts (*mortality_rates_educ.do* and *import_seer_population.R*, described in Section 4.1), which are not called by any reports and only clean data, are also located in this folder.

5 HRS pipeline

Some minor analyses in this paper require the use of restricted-use Health and Retirement Study (HRS) data, which can only be accessed within the physical NBER building at 1050 Massachusetts Avenue. We created a small auxiliary pipeline for the sole purpose of analyzing these data, and copied a version of this pipeline (without the restricted-use data) to the "HRS" folder in the replication kit. Below, we share a quick guide for interpreting this data.

5.1 Data

Data is organized relatively similarly to the NCHS and Medicare pipeline. A "raw" folder contains three files:

- *state_population_SEER*: State-level population statistics that are equivalent to those used across other analyses.
- *yagan_replication_files*: Replication files from Yagan (2019)[4] which contain estimates of state-level unemployment shock.
- *Public_HRS2018*: Public-access HRS data which complements restricted-access data.

When conducting actual analyses within the NBER, there is a fourth, restricted-access file which we removed before constructing this replication kit.

5.2 Analysis

Note that there is no "build" or "clean_data" folder in this abridged pipeline, as there are no intermediate data cleaning steps required. Instead, all analysis is done within *code/cross_section_analysis.do*. Only four figures are produced in this program, all of which are produced in the output folder. These four figures correspond to the four panels of Figure OA.31.

6 Appendix

6.1 Acronyms

- ACS: American Community Survey, runs annually and collects information on age, education, location
- BLS: Bureau of Labor Statistics
- BRFSS: Behavior Risk Factor Surveillance Survey, tracks morbidity and health behaviors for the surveyed population
- CDC: Centers for Disease Control and Prevention; source for much of our mortality data
- CPI: Consumer Price Index
- CPS: Current Population Survey; run by the US Census and source of some earnings data for the paper
- CZ: Commuting zone; frequent unit of geographic analysis throughout the paper
- EPOP: Employment-population ratio
- GR: Great Recession
- HRS: Health Research Survey, a survey on individuals' health behaviors for which the restricted version can only be analyzed at the NBER
- HPI: House Price Index
- IPUMS: Integrated Public Use Microdata Series; contains Census and survey data used in the paper
- LAUS: Local Area Unemployment Statistics, produced by the Bureau of Labor Statistics. We use this data for most of our labor force participation statistics, including unemployment shocks.

- MA: Medicare Advantage
- NBER: National Bureau of Economic Research
- NCHS: National Center for Health Statistics; source for much of our mortality data
- OSCAR: Online Survey Certification and Reporting; database used to examine SNF-level characteristics
- SEER: Surveillance Epidemiology and End Results, run by the National Cancer Institute; this is the source for much of our population data throughout the paper
- SNF: Skilled nursing facility
- SSA: Social Security Administration, produces life tables used for summary statistics
- TM: Traditional Medicare

References

[1] David Autor, David Dorn, and Gordon Hanson. The China Shock: Learning from Labor-Market Adjustment to Large Changes in Trade. *Annual Reviews*, 8:205–240, 2016.

[2] Alexander W. Bartik, Janet Currie, Michael Greenstone, and Christopher R. Knittel. The Local Economic and Welfare Consequences of Hydraulic Fracturing. *American Economic Journal: Applied Economics*, 11(4):105–155, October 2019.

[3] Liyang Sun and Jesse M. Shapiro. A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure. *Journal of Economic Perspectives*, 36(4):193–204, November 2022.

[4] Danny Yagan. Employment Hysteresis from the Great Recession. *Journal of Political Economy*, 127(5):2505–2558, 2019. Number: 5.