

Pulling and processing the data

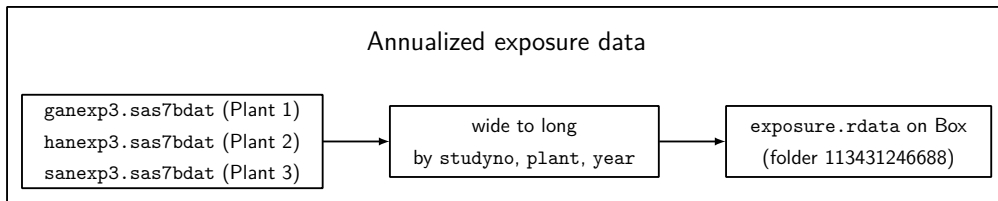
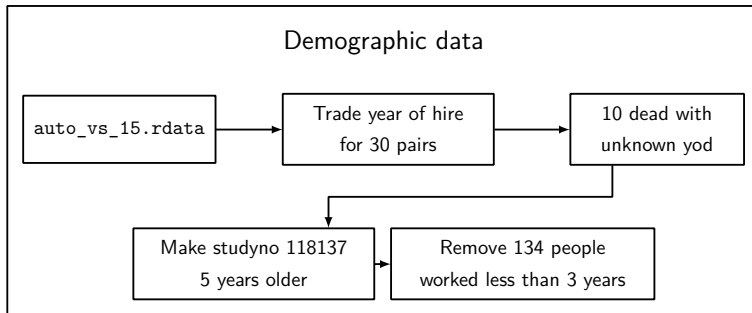
GM-UAW Cohort Study

January 10, 2021

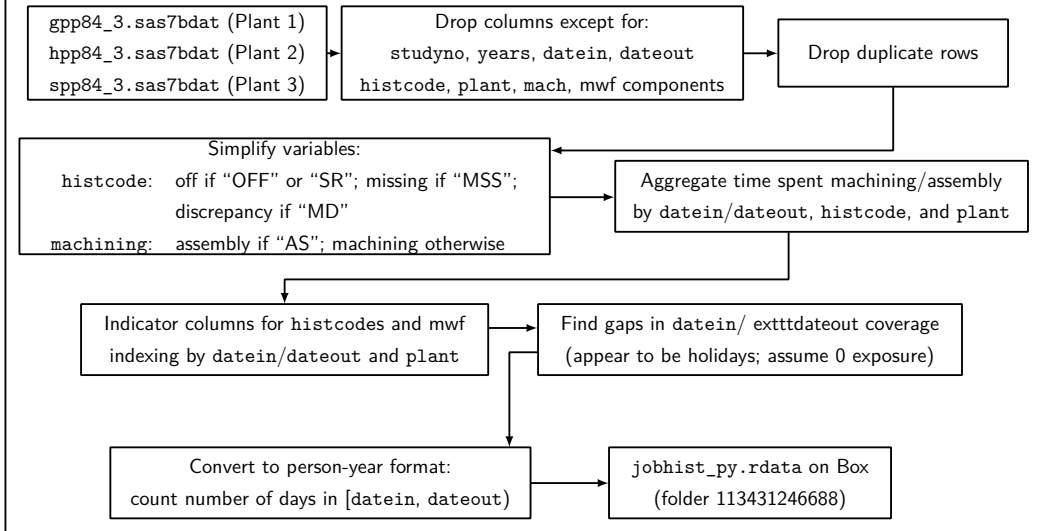
Overview

- ▶ Pulling and preliminary cleaning
- ▶ Building analytic dataset
 - ▶ Demographic data in person-time format
 - ▶ Map ICD codes to causes of death/cancer types
 - ▶ Merge exposure data and job history data
- ▶ Intermediate objects are saved in Box folder 113431246688
- ▶ R code for doing all of this on Github

Pulling and preliminary cleaning



Job history data



Person-year dataset

- ▶ Start with the cleaned demographic data cohort
- ▶ Duplicate each row so that each year from hire to death/end of FU is represented
- ▶ Index appropriately by calendar year and age

Outcome labels

- ▶ Cause of death labels coded up using ICD mappings given by the NIOSH-92 death categories document (linked here)
- ▶ Cancer incidence from MCR coded up by Liza (thank you!)
- ▶ Cancer incidence from SEER coded up using the Site Recode ICD-O-3/WHO 2008 Definitions (linked here), taking into account both the ICD-O-3 Site code *and* the ICD-O-3 Histology code
 - ▶ Cleaned up SEER incidence data frames on Box (see `SEER_incidence.csv` in (folder 113431246688)[<https://berkeley.app.box.com/folder/113431246688>])

Merging exposure and job history data

- ▶ Recall that exposure data was indexed by `studyno`, `year`, *and* `plant`
 - ▶ Before merging, exposure was summed across plants i.e. indexed by `studyno` and `year` only
- ▶ `plant` was taken to be plant with the most days in `jobhist_py.rdata`, for that `year`

Code: it's all on github

The screenshot shows the GitHub interface for the repository `tao-feng / gm-wrangling`. The top navigation bar includes links for Pull requests, Issues, Marketplace, and Explore. The repository header shows 1 pull request, 0 stars, and 0 forks. The main content area has tabs for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, and Insights. The 'Code' tab is active, displaying a file tree with folders like 'cancer incidence', 'causes of death', 'docs', and 'wrangling', and files like 'cancer-key.tsv' and 'readme.md'. The 'About' section on the right provides information about the repository, including a README link, release status, package status, and languages.

Search or jump to... Pull requests Issues Marketplace Explore

tao-feng / gm-wrangling Unwatch 1 Star 0 Fork 0

<> Code Issues Pull requests Actions Projects Wiki Security Insights

master 1 branch 0 tags Go to file Add file Code

tao-feng sync 3a095ae 23 days ago 15 commits

folder	cancer incidence	Sync up	6 months ago
folder	causes of death	sync	23 days ago
folder	docs	some documentation for get.cohort_analytic	last month
folder	wrangling	some documentation for get.cohort_analytic	last month
file	cancer-key.tsv	first commit	6 months ago
file	readme.md	some documentation for get.cohort_analytic	last month

readme.md

title	subtitle	author	date
-------	----------	--------	------

About Get and reshape UAW-GM.

Readme

Releases No releases published [Create a new release](#)

Packages No packages published [Publish your first package](#)

Languages

Getting the code

```
#!/bin/sh

cd "~" # Must must clone into your home directory
git clone https://github.com/tao-feng/HeadRs.git # Dependencies

cd "directory/of your/choice"
git clone https://github.com/tao-feng/gm-wrangling.git
```

Or Download ZIP – after unzipping, please change the directory names to “HeadRs” and “gm-wrangling”

The home directory R sees can be found by running `path.expand("~")`

Getting the data and helper functions

```
# Check that necessary packages are installed  
lapply(c("tidyverse", "xtable", "pander", "tikzDevice", "knitr",  
        "data.table", "zoo", "boxr", "lubridate", "sas7bdat", "Hmisc"),  
       function(package) {if (!package %in% installed.packages()) {  
         install.packages(package)}})  
  
# Get data and helper functions  
source("directory/of your/choice/gm-wrangling/wrangling/00-hello.R")
```

Note: Setting up boxr is a bit of a pain...

What you get from running 00-hello.R

additional_outcomes()	get.cohort_py()	jobhist	mytheme.web
cohort	get.exposure()	jobhist_py	og.dir
date.to.gm()	get.jobhist()	jobhist_py.cast	self_injury.function()
death_type()	get.ltab_obs()	latex()	spec_icd_codes()
drive_D	gm.to.date()	ltab_age()	tikzLualatexPackages.option
dta	hook_output()	ltab_calendar()	to_drive_D()
exposure	icd_codes.function()	lualatex()	
get.cohort_analytic()	is.auto_vs_15	mytheme	

Notes

- ▶ cohort is the cleaned demographic data
 - ▶ Please check variable names and types
 - ▶ Most variables correspond to those in auto_vs_15
- ▶ exposure is the pulled/merged exposure data
- ▶ jobhist_py is the cleaned person-year job history data

Making the analytic data object