Pulling and processing the data

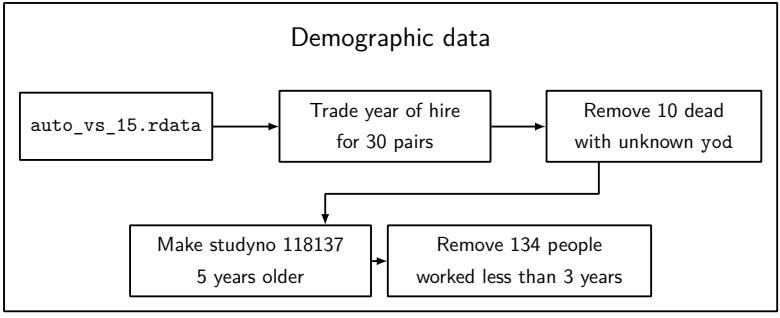
UAW-GM Cohort Study

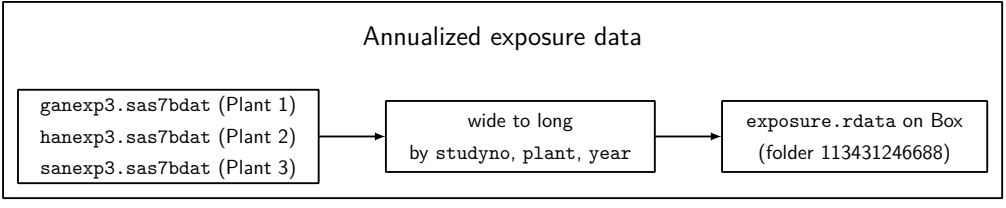
January 13, 2021

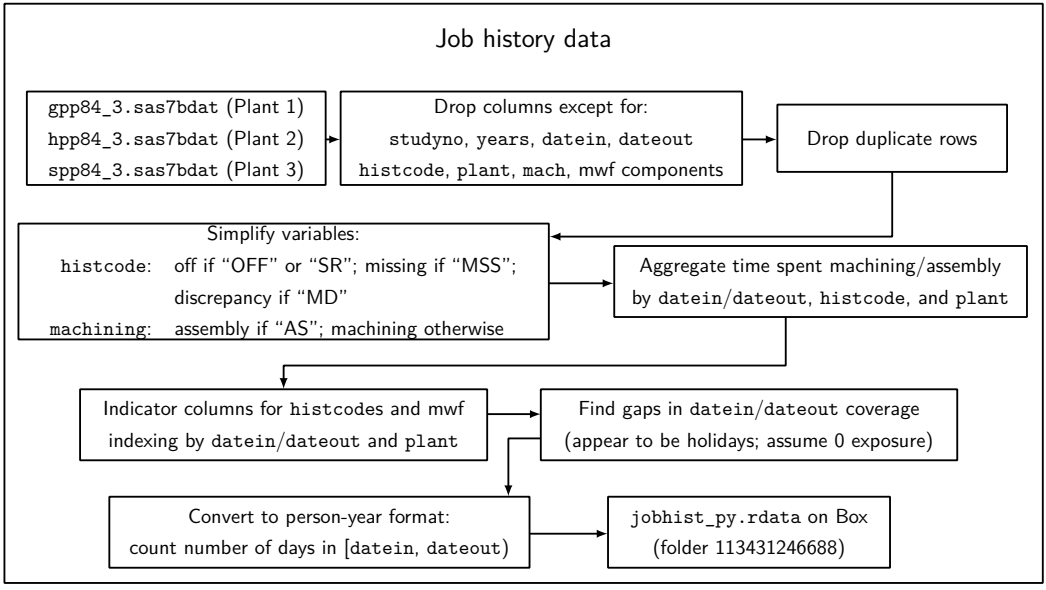
Overview

- 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
gpp84_3.sas7bdat (Plant 1)
hpp84_3.sas7bdat (Plant 2)
spp84_3.sas7bdat (Plant 3)
Drop columns except for:
studyno, years, datein, dateout
histcode, plant, mach, mwf components
Drop duplicate rows
Simplify variables:
histcode: off if "OFF" or "SR"; missing if "MSS";
discrepancy if "MD"
machining: assembly if "AS"; machining otherwise
Aggregate time spent machining/assembly
by datein/dateout, histcode, and plant
Indicator columns for histcodes and mwf
indexing by datein/dateout and plant
Find gaps in datein/dateout coverage
(appear to be holidays; assume 0 exposure)
Convert to person-year format:
count number of days in [datein, dateout)
jobhist_py.rdata on Box
(folder 113431246688)

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

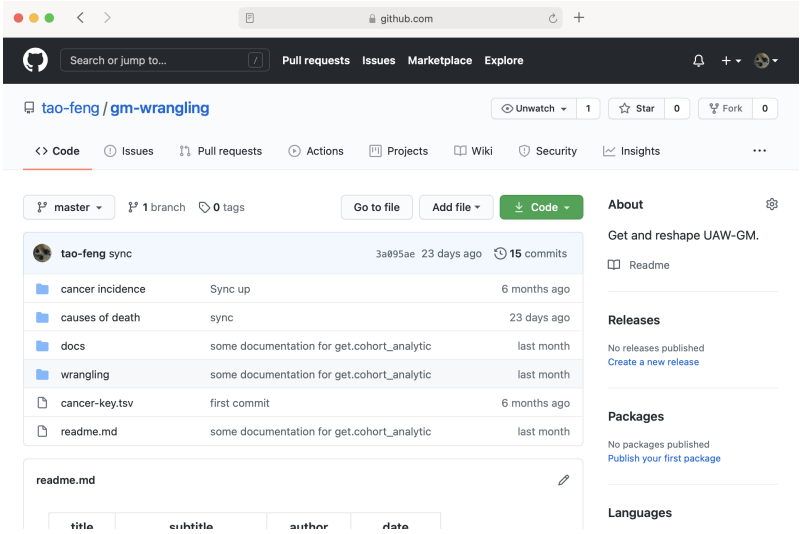
Outcome labels

- 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)

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



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

- Running `00-hello.R` performs all the tasks outlined in the flow charts above
- `get.cohort_py()` makes the person-time dataset
- `get.exposure()` merges the exposure data
- `get.ltab_obs()` runs `get.cohort_py()` and `get.exposure()` to generate a person-time dataset with the demographic data and cause of death indicator columns
- `get.cohort_analytic()` runs `get.ltab_obs()` to generate an analytic dataset with the demographic data, exposure data, and indicator columns for different causes of death and incident cancers

Example of use

For the cancer mortality paper Costello et al. (2020), the following works

```
# Get data and helper functions
source("directory/of your/choice/gm-wrangling/wrangling/00-hello.R")
# Get data
cohort_analytic <- get.cohort_analytic(
        outcome_type = "mortality",
        exposure.lag = 21,
        deathage.max = NULL)
# Filter data
cohort_analytic <- cohort_analytic[year >= 1941 & (
        year(yin) < 1938 | year >= year(yin + 365.25 * 3))]
```

Example of use (continued)

studyno	year	age.year1	age.year2	All causes
100001	1974	11727	12053	0
100001	1975	12053	12418	0
100001	1976	12418	12784	0
100001	1977	12784	13149	0
⋮	⋮	⋮	⋮	⋮
100001	2013	25933	26298	0
100001	2014	26298	26318	1