

Cancer incidence analytic dataset

UAW-GM Cohort Study

Kevin Chen

This document briefly describes the data saved in `Box/Autoworkrs/Kevin/Cancer Incidence (2020)/data`. Relevant github repositories are (1) `HeadRs`, which includes script defining helper functions, L^AT_EX headers, and assorted style files; (2) `gm-wrangling`, which includes script for loading and cleaning UAW-GM Cohort data (requires authentication using `boxr`); and (3) `gm-cancer-inc`, which contains materials specific to the cancer incidence work presented in the parent of the current directory.

The data saved here was generated by running the chunk below. Please have repositories `HeadRs` and `gm-wrangling` cloned into the parent directory of the working directory.

```
# R version 4.0.0 (2020-04-24)
# Platform: x86_64-apple-darwin17.0 (64-bit)

library(here)
library(boxr); box_auth()
source(here::here('../gm-wrangling/wrangling', '00-hello.R'))
yout.which <- "yout"

cohort_analytic <- get.cohort_analytic(
  outcome_type = "incidence",
  exposure.lag = 21,
  deathage.max = NULL,
  end.year = 2015,
  hire.year.min = -Inf,
  use_seer = T
)
setorder(cohort_analytic, studyno, year)
cohort_analytic[, `:=`(yin.gm = date.to.gm(yin))]

# Keep only people who appear in the exposure data
cohort_analytic <-
  cohort_analytic[studyno %in% unique(exposure$studyno)]

# PICK YOUT ####
cohort_analytic[, jobloss.date := get(yout.which)]

# Exposure after leaving work is 0
```

```

cohort_analytic[year > (year(jobloss.date) + exposure.lag), `:=`(
  straight = 0,
  soluble = 0,
  synthetic = 0)]
# NA fill
cohort_analytic[year <= (year(jobloss.date) + exposure.lag), `:=`(
  straight = zoo::na.locf(straight),
  soluble = zoo::na.locf(soluble),
  synthetic = zoo::na.locf(synthetic)
), by = .(studyno)]
cohort_analytic[, `:=`(
  cum_straight = cumsum(straight),
  cum_soluble = cumsum(soluble),
  cum_synthetic = cumsum(synthetic)
), by = .(studyno)]

# Upload to Box (about 60.3 MB)
box_save(cohort_analytic,
  dir_id = 118903632077,
  file_name = "cohort_analytic.rdata",
  description = paste0(
    "Analytic data for cancer incidence analyses ",
    "of class `data.table` (dimensions: ",
    paste(dim(cohort_analytic), collapse = " x "), ")")

# Simplify data structure
cohort_analytic[, `:=`(
  canc_which_first = sapply(canc_which_first, function(x) {
    paste(x, collapse = ", ")
  })
)]

# Upload to Box (about 1.7 GB)
box_write(cohort_analytic,
  dir_id = 118903632077,
  file_name = "cohort_analytic.csv",
  description = paste0(
    "Analytic data for cancer incidence analyses (dimensions: ",
    paste(dim(cohort_analytic), collapse = " x "), ")")

```

In the analyses presented in `../reports`, the data were restricted to rows satisfying all of the following: `wh == 1`, `nohist == 0`, `possdiscr_new == 0`, `immortal == 0`, `right.censored == 0`. Note that the data in the current directory were not filtered in this way. Before passing the data to functions for analysis, it would be advisable to check these flags as well as the start/end times for each row. The time indexing variable pairs `age.year1/age.year2` and `year1/year2` attempt to follow the indexing conventions of the survival package (see vignette on time-dependent variables).

The table below provides descriptions of the variables found in the uploaded data. Documentation for most of these variables exist elsewhere on Box.

Variable	Description
<code>studyno</code>	Unique identifier
<code>age.year1</code>	Age at the beginning of the at-risk person-year (3 years after hire or 1973-01-01 for plants 1 and 2 or 1985-01-01 for plant 3, whichever is later) in days
<code>age.year2</code>	Age at the end of the at-risk person-year (death or censoring or 2015-12-31, whichever is earlier) in days
<code>year1</code>	Date of the start of the person-year at-risk
<code>year2</code>	Date of the end of the person-year at-risk
<code>canc_**</code>	Cancer incidence indicator (includes both SEER and MCR follow-up). A value of 1 indicates cancer incidence in that year. A value of 2 indicates cancer incidence in previous years. A value of 0 indicates no recorded cancer incidence up to the end of that year. The asterisks are placeholder for cancer type abbreviations. See sheet 2 of the data dictionary on Box.
<code>canc_first</code>	Cancer incidence indicator for first instance of any cancer type, coded in the same way as the other <code>canc_**</code>
<code>canc_which_first</code>	Comma-delimited codes denoting the type(s) of the first cancer (up to 7 different types on the same day)
<code>straight</code>	Exposure to straight metalworking fluids (mg/m^3) lagged 21 years
<code>cum_straight</code>	Cumulative exposure to straight metalworking fluids lagged 21 years ($\text{mg}/\text{m}^3 \cdot \text{years}$)
<code>soluble</code>	Exposure to soluble metalworking fluids (mg/m^3) lagged 21 years
<code>cum_soluble</code>	Cumulative exposure to soluble metalworking fluids lagged 21 years ($\text{mg}/\text{m}^3 \cdot \text{years}$)
<code>synthetic</code>	Exposure to synthetic metalworking fluids (mg/m^3) lagged 21 years
<code>cum_synthetic</code>	Cumulative exposure to synthetic metalworking fluids ($\text{mg}/\text{m}^3 \cdot \text{years}$) lagged 21 years
<code>bio.**</code>	Proportion of the calendar year exposed to biocides (asterisks in place of site <code>gan</code> plant 1, <code>han</code> plant 2, or <code>san</code> plant 3)
<code>cl.**</code>	Proportion of the calendar year exposed to chlorine
<code>ea.**</code>	Proportion of the calendar year exposed to ethanolamine
<code>tea.**</code>	Proportion of the calendar year exposed to triethanolamine
<code>trz.**</code>	Proportion of the calendar year exposed to triazine
<code>s.**</code>	Proportion of the calendar year exposed to sulfur
<code>no2.**</code>	Proportion of the calendar year exposed to nitrites
<code>year</code>	Calendar year
<code>yin.gm</code>	Date of hire as a floating point number
<code>yin</code>	Date of hire

Variable	Description
<code>yrin</code>	Date of start of at-risk time for mortality follow-up i.e. 3 years after hire
<code>yrin16</code> ‘	Date of start of at-risk time for mortality follow-up as a floating point number
<code>race</code>	Race (those with unknown race coded as White)
<code>finrace</code>	Race as coded in the original data (See description for FINRACE in the data dictionary on Box)
<code>plant</code>	Plant (time-varying)
<code>ddiag_**</code>	Date of cancer incidence, if exists, with asterisks in place of cancer type codes
<code>ddiag_first</code>	Date of first cancer incidence, if exists
<code>yod</code>	Date of death, if exists
<code>yoc</code>	Date of censoring (upon reaching the oldest observed age at death: 108.39 years)
<code>yob</code>	Date of birth
<code>sex</code>	Sex: “M” or “F”
<code>dateout.date</code>	Last date out in job history records
<code>employment_end.date</code>	End of employment date (made for the suicide and job loss analyses)
<code>employment_end.date.legacy</code>	End of employment date (an earlier version of employment_end.date)
<code>yout</code>	End of employment date as in the original data (see description for YOUT16 the data dictionary on Box)
<code>yout_recode</code>	End of employment date reconstructed from old documentation
All causes	Indicator of death due to any cause. A value of 1 indicates death in that year, 0 otherwise.
Chronic obstructive pulmonary disease	Indicator of death due to chronic obstructive pulmonary disease. A value of 1 indicates death due to COPD in that year, 0 otherwise.
All external causes	Indicator of death due to external causes. A value of 1 indicates death due to external causes in that year, 0 otherwise.
<code>nohist</code>	Indicator of whether individual appears in the job history records (see description for NOHIST the data dictionary on Box)
<code>wh</code>	Indicator of whether an individual with at least one job history record also has at least half of their work record not missing (see description for WH the data dictionary on Box)
<code>immortal</code>	Indicator of immortal person-year
<code>right.censored</code>	Indicator of person years considered lost to follow-up
<code>possdiscr_new</code>	Indicator of possible discrepancy in matching/merging (see description for POSSDISCR_NEW the data dictionary on Box)
<code>flag77</code>	Indicator of possible issue involving date of leaving work (see description for flag77 the data dictionary on Box)

Variable	Description
<code>oddend</code>	Indicator of possible issue involving the job history data (see description for <code>oddend</code> the data dictionary on Box)
<code>status15</code>	Vital status through 2015. Values 3, 6, and 9 indicate alive, dead, and unknown, respectively (see description for <code>STATUS15</code> the data dictionary on Box)
<code>cancinccoh15_new</code>	Indicator of eligibility for cancer incidence follow-up (see description for <code>cancinccoh15_new</code> the data dictionary on Box)