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, LaTeX 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 as rdata (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 = ", ")
    })
)]
cohort_analytic <- as.data.frame(cohort_analytic)

# Upload to Box as csv (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 "), ")"))

# Sanitize names
names(cohort_analytic) <- gsub("-|\\.|\\/| ", "_", names(cohort_analytic))
names(cohort_analytic) <- gsub(",|\\'|\\[|\\]|\\(|\\)", "", names(cohort_analytic))
names(cohort_analytic)[sapply(names(cohort_analytic), nchar) > 32] <- c(
    "Accidents",
    "b_duct_liver_g_bladder_cancers",
```

```
    "bladder_urinary_cancers",
    "COPD",
    "Chronic_liver_disease",
    "nonmalignant_respiratory_dis"
)
# Upload to Box as xpt (about 4.1 GB)
box_write(x = cohort_analytic,
                    dir_id = 118903632077,
                    file_name = "cohort_analytic.xpt",
                    write_fun = haven::write_xpt,
                    description = paste0(
                        "Analytic data for cancer incidence analyses (dimensions: ",
                        paste(dim(cohort_analytic), collapse = " x "), ")"),
                    .name_repair = "universal")

# Upload to Box as sas7bdat (about 5.4 GB)
box_write(x = cohort_analytic,
                    dir_id = 118903632077,
                    file_name = "cohort_analytic.sas7bdat",
                    write_fun = haven::write_sas,
                    description = paste0(
                        "Analytic data for cancer incidence analyses (dimensions: ",
                        paste(dim(cohort_analytic), collapse = " x "), ")"),
                    .name_repair = "universal")
```

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).

Table 1: The table below provides descriptions of the variables found in the uploaded data. Documentation for most of these variables exist elsewhere on Box. Variable names in `SAS`-formatted files have been sanitized an do not appear as they do here.

Variable	Description
`studyno`	Unique identifier
`age.year1`	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
`age.year2`	Age at the end of the at-risk person-year (death or censoring or 2015-12-31, whichever is earlier) in days
`year1`	Date of the start of the person-year at-risk
`year2`	Date of the end of the person-year at-risk

Variable	Description
canc_**	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.
canc_first	Cancer incidence indicator for first instance of any cancer type, coded in the same way as the other canc_**
canc_which_first	Comma-delimited codes denoting the type(s) of the first cancer (up to 7 different types on the same day)
straight	Exposure to straight metalworking fluids (mg/m^3) lagged 21 years
cum_straight	Cumulative exposure to straight metalworking fluids lagged 21 years (mg/m^3·years)
soluble	Exposure to soluble metalworking fluids (mg/m^3) lagged 21 years
cum_soluble	Cumulative exposure to soluble metalworking fluids lagged 21 years (mg/m^3·years)
synthetic	Exposure to synthetic metalworking fluids (mg/m^3) lagged 21 years
cum_synthetic	Cumulative exposure to synthetic metalworking fluids (mg/m^3·years) lagged 21 years
bio.**	Proportion of the calendar year exposed to biocides (asterisks in place of site gan plant 1, han plant 2, or san plant 3)
cl.**	Proportion of the calendar year exposed to chlorine
ea.**	Proportion of the calendar year exposed to ethanolamine
tea.**	Proportion of the calendar year exposed to triethanolamine
trz.**	Proportion of the calendar year exposed to triazine
s.**	Proportion of the calendar year exposed to sulfur
no2.**	Proportion of the calendar year exposed to nitrites
year	Calendar year
yin.gm	Date of hire as a floating point number
yin	Date of hire
yrin	Date of start of at-risk time for mortality follow-up i.e. 3 years after hire
yrin16 ‘	Date of start of at-risk time for mortality follow-up as a floating point number
race	Race (those with unknown race coded as White)
finrace	Race as coded in the original data (See description for FINRACE in the data dictionary on Box)
plant	Plant (time-varying)
ddiag_**	Date of cancer incidence, if exists, with asterisks in place of cancer type codes
ddiag_first	Date of first cancer incidence, if exists

Variable	Description
yod	Date of death, if exists
yoc	Date of censoring (upon reaching the oldest observed age at death: 108.39 years)
yob	Date of birth
sex	Sex: "M" or "F"
dateout.date	Last date out in job history records
employment_end.date	End of employment date (made for the suicide and job loss analyses)
employment_end.date.legacy	End of employment date (an earlier version of employment_end.date)
yout	End of employment date as in the original data (see description for YOUT16 the data dictionary on Box)
yout_recode	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.
nohist	Indicator of whether individual appears in the job history records (see description for NOHIST the data dictionary on Box)
wh	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)
immortal	Indicator of immortal person-year
right.censored	Indicator of person years considered lost to follow-up
possdiscr_new	Indicator of possible discrepancy in matching/merging (see description for POSSDISCR_NEW the data dictionary on Box)
flag77	Indicator of possible issue involving date of leaving work (see description for flag77 the data dictionary on Box)
oddend	Indicator of possible issue involving the job history data (see description for oddend the data dictionary on Box)
status15	Vital status through 2015. Values 3, 6, and 9 indicate alive, dead, and unknown, respectively (see description for STATUS15 the data dictionary on Box)
cancinccoh15_new	Indicator of eligibility for cancer incidence follow-up (see description for cancinccoh15_new the data dictionary on Box)