Join NHANES data with linked NDI mortality data

Martin Skarzynski

2018-04-19

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## read in data processed using sas ####
adult <- readr::read_csv(here::here("dat/adult.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   DMPSTAT = col_integer(),
##   DMARETHN = col_integer(),
##   DMARACER = col_integer(),
##   DMAETHNR = col_integer(),
##   HSSEX = col_integer(),
##   HSAGEIR = col_integer(),
##   HSAGEU = col_integer(),
##   DMPMETRO = col_integer(),
##   DMPCREGN = col_integer(),
##   SDPPHASE = col_integer(),
##   SDPPSU6 = col_integer(),
##   SDPPSU1 = col_integer(),
##   MXPLANG = col_integer(),
##   MXPSESSR = col_integer(),
##   MXPTIDW = col_integer(),
##   HXPSESSR = col_integer(),
##   HXPTIDW = col_integer(),
##   HFVERS = col_integer(),
##   HFINTVR = col_integer(),
##   HFLANG = col_integer()
##   # ... with 565 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
mort <- readr::read_rds(here::here("dat/1-clean-mort.rds"))
exam <- readr::read_csv(here::here("dat/exam.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
```

```
##   DMPPIR = col_character(),
##   SDPPSU2 = col_character(),
##   SDPSTRA2 = col_character(),
##   WTPFQX6 = col_character(),
##   WTPFEX6 = col_character(),
##   WTPFHX6 = col_character(),
##   WTPFALG6 = col_character(),
##   WTPFCNS6 = col_character(),
##   WTPFSD6 = col_character(),
##   WTPFMD6 = col_character(),
##   WTPFHSD6 = col_character(),
##   WTPFHMD6 = col_character(),
##   WTPFQX1 = col_character(),
##   WTPFEX1 = col_character(),
##   WTPFHX1 = col_character(),
##   WTPFALG1 = col_character(),
##   WTPFCNS1 = col_character(),
##   WTPFSD1 = col_character(),
##   WTPFMD1 = col_character(),
##   WTPFHSD1 = col_character()
##   # ... with 648 more columns
## )
## See spec(...) for full column specifications.
```

```
lab <- readr::read_csv(here::here("dat/lab.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   SEQN = col_integer(),
##   DMPFSEQ = col_integer(),
##   DMPSTAT = col_integer(),
##   DMARETHN = col_integer(),
##   DMARACER = col_integer(),
##   DMAETHNR = col_integer(),
##   HSSEX = col_integer(),
##   HSAGEIR = col_integer(),
##   HSAGEU = col_integer(),
##   HSAITMOR = col_integer(),
##   HSFSIZER = col_integer(),
##   HSHSIZER = col_integer(),
##   DMPCNTYR = col_integer(),
##   DMPFIPSR = col_integer(),
##   DMPMETRO = col_integer(),
##   DMPCREGN = col_integer(),
##   SDPPHASE = col_integer(),
##   SDPPSU6 = col_integer(),
##   SDPSTRA6 = col_integer(),
##   SDPPSU1 = col_integer()
##   # ... with 190 more columns
## )
## See spec(...) for full column specifications.
```

```
## change SEQN to numeric in all datasets read in from csv
adult$SEQN <- as.numeric(adult$SEQN)
```

```r
exam$SEQN <- as.numeric(exam$SEQN)
lab$SEQN <- as.numeric(lab$SEQN)

#levels(dat$UCOD_LEADING)

## Join all datasets, remove baseline cancer cases
## Create a cancer death variable
## Convert all character variables to numeric
## Select only columns with less than 10% NAs
## write out an RDS file
left_join(x = mort, y = adult, by = "SEQN") %>%
left_join(x = ., y = exam, by = "SEQN") %>%
left_join(x = ., y = lab, by = "SEQN") %>%
filter(HAC1N==2 &
       HAC1O==2 &
       !is.na(SDPPSU6) &
       !is.na(SDPSTRA6) &
       !is.na(WTPFQX6)) %>%
mutate(canc_mort =
       if_else(UCOD_LEADING ==
               'Malignant neoplasms (C00-C97)',
               true = 1, false = 0)) %>%
mutate_if(.predicate = is.character,
          .funs = as.numeric) %>%
select(which(colMeans(is.na(.))==0)) %>%
readr::write_rds(here::here("dat/2-join-complete-cases.rds"))
```

```
## Warning in evalq(as.numeric(HAJ12), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(HAX18A), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(HAX18B), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(HAX18C), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(HAZA1CC), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(DEPSTLC1), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(DEPSTLC2), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(DEPSTLC3), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(DEPSTLC4), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(DEPSTLC5), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(DEPSTLC6), <environment>): NAs introduced by
## coercion
```

```
## Warning in evalq(as.numeric(SPPTIME), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(PHPSNTI), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(PHPDRTI), <environment>): NAs introduced by
## coercion

## Warning in evalq(as.numeric(PHPBEST), <environment>): NAs introduced by
## coercion
```

```
#warnings()
## 15 variables (HAJ12, HAX18A, HAX18B, HAX18C, HAZA1CC, DEPSTLC1, DEPSTLC2, DEPSTLC3, DEPSTLC4, DEPSTL
## 3 time variables coerced into NA (PHPSNTI, PHPDRTI, PHPBEST)
## Consider converting these 3 to a different class later
```