Exploratory data analysis of the HRS cohort

Peter T. Tanksley, Ph.D.

Tentative article title: “Evolutionary gerontology and the study of criminal justice health disparities: A genetic analysis of Alzheimer’s Disease in the Health and Retirement Study”

This is an exploratory exercise to identify the characteristics of the analytic sample we will be using from the HRS cohort. The data come from several sources and has been subjected to only minimal data cleaning prior to being merged for this EDA. It is contains repeated observations.

```r
#load eda packages
library(pacman)
p_load(rio,
       tidyverse,
       # tidylog,
       ###
       naniar,
       sjlabelled,
       skimr)

#read in merged hrs data (long format)
hrs <- import("hrs_merged.rds")

#dataframe dimensions (rows, columns)
dim(hrs)
```

```
[1] 777116     27
```

So far everything is in good order. Let’s see how many observations vs unique cases we have in the data.

```r
hrs %>%
  count(cases = n_distinct(hhidpn))
```

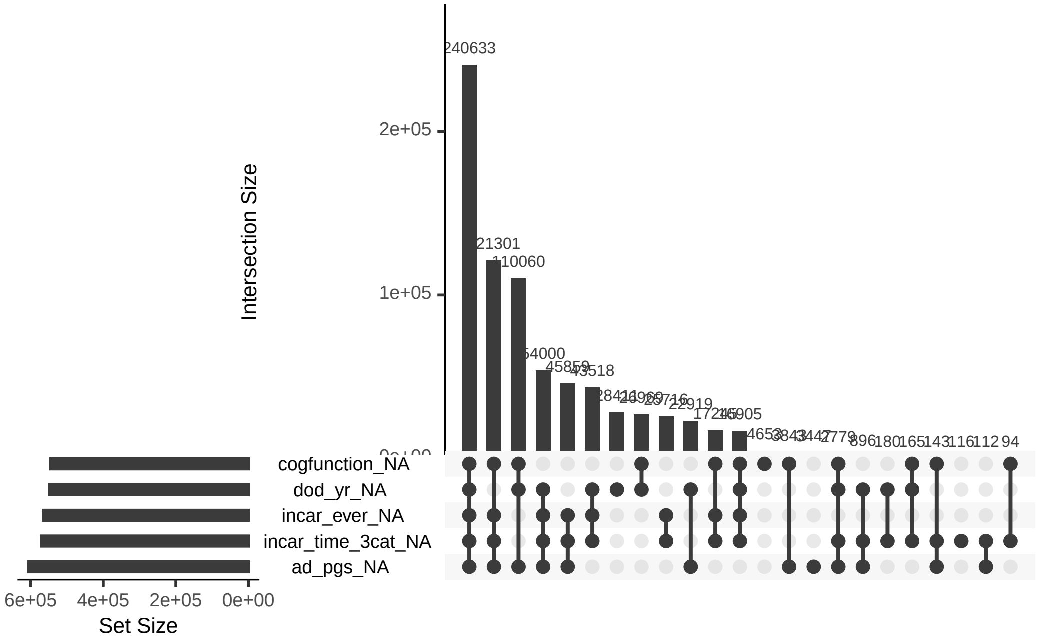
```
# A tibble: 1 x 2
  cases      n
  <int>  <int>
1 57205 777116
```

So we have ~57K individuals participating in the HRS between 1992-2018, with ~777K observations between them. We have a couple of factors that will whittle these numbers down though:

- How many individuals contributed DNA?
  - We know this number to be around 20K from the documentation, and only around 17K of those are EUR ancestry (the group we will focus on)
- How many individuals provided incarceration data?
  - In 2012/2014, participants we provided with a leave-behind questionnaire (LBQ) in which were the items we'll use. Not everyone returned the LBQ, however, again limiting our sample (this time to those people to went the extra mile–selection effects???).
- Additionally, alot of people died during the study (the HRS is comprised of elderly Americans after all). So attrition from death is another issue.

Let's take a look at the data without any restrictions and see what our missingness looks like.

```r
hrs %>%
  select(-starts_with("pc")) %>% #we don't need to see ancestry PCs
  naniar::gg_miss_upset()
```

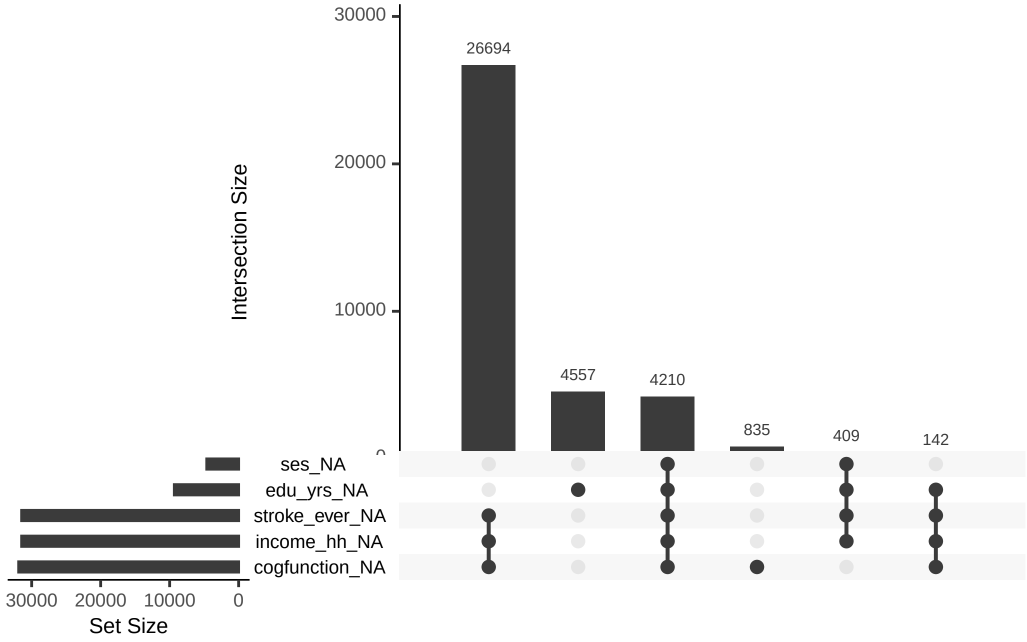


Ok, we are missing alot of information for people on three key variables (cognitive function, Alzheimer's disease [AD] PGS, incarceration) and our item for year of death. That last one is ok though because if a person wasn't reported as "deceased" in the most recent wave of data collection then they wouldn't have a value reported (because they're still kicking).

Let's try filtering things a bit and see what the picture is. We'll filter on the following:

- Death prior to 2012 (i.e., when incarceration data were first collected)
- AD PGS (this was calculated for all participants with valid DNA data)
- Incarceration (only collected for two waves [2012/2014], but any reported value is copied forward/backward)

```
hrs %>%
  select(-starts_with("pc")) %>% #we still don't need to see ancestry PCs
  filter(as_numeric(dod_yr) >= 2012 | is.na(dod_yr)) %>% #removed 156,043 rows (20%), 621,
  filter(!is.na(ad_pgs)) %>% #465,438 rows (75%), 155,635 rows remaining
  filter(!is.na(incar_ever)) %>% #87,895 rows (56%), 67,740 rows remaining
  select(-dod_yr) %>%
  naniar::gg_miss_upset()
```



Our sample went way down (from 750K observations to ~68K), but we expected that. The remaining variables that we're missing data on cognitive function and some key covariates: household income, and history of stroke. Let's see what we're left with if we do an across the board listwise deletion.

```r
hrs %>%
  filter(as_numeric(dod_yr) >= 2012 | is.na(dod_yr)) %>%
  select(hhidpn,
         study, race_ethn, sex, birthyr,
         cogfunction,
         ad_pgs, starts_with("pc"),
         incar_ever,
         stroke_ever,
         apoe_info99_4ct,
         social_origins,
         ses
         # edu_yrs,
         # income_hh
         ) %>%
  drop_na() %>%
  count(cases = n_distinct(hhidpn))
```

```
# A tibble: 1 x 2
  cases     n
  <int> <int>
1  4428 34772
```

This process has given us a better idea of the sample size we're working with. Unfortunately, the particular intersection of (1) people who participate in population health studies, (2) have a CJ background, and (3) are willing to provide DNA, makes our study population fairly small.

But onward and upward! Let's see some descriptive statistics.

```
#construct our study sample (including only individuals who did not die prior 2012)
hrs %>%
  filter(as_numeric(dod_yr) >= 2012 | is.na(dod_yr)) %>%
  select(hhidpn,
         study, race_ethn, sex, birthyr,
         cogfunction,
         ad_pgs, #starts_with("pc"),
         incar_ever,
         stroke_ever,
         apoe_info99_4ct,
         social_origins,
         ses) %>%
  drop_na() %>%
  skim()
```

Table 1: Data summary

Name	Piped data
Number of rows	34772
Number of columns	12
Column type frequency:	
factor	7
numeric	5
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
hhidpn	0	1	FALSE	4428	110: 11, 110: 11, 110: 11, 110: 11
study	0	1	FALSE	5	WB: 10303, EBB: 9319, COD: 5535, MBB: 5383
race_ethn	0	1	FALSE	1	Whi: 34772, Bla: 0, His: 0, Oth: 0
sex	0	1	FALSE	2	Fem: 20596, Mal: 14176
cogfunction	0	1	FALSE	3	nor: 31101, cin: 3034, dem: 637
incar_ever	0	1	FALSE	2	Not: 32313, Inc: 2459
apoe_info99_4ct	0	1	FALSE	3	zer: 26654, one: 7452, two: 666

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
birthyr	0	1	1943.86	12.40	1911.00	1930.00	1947.00	1953.00	1980.00	
ad_pgs	0	1	-0.05	1.01	-3.49	-0.75	-0.07	0.63	3.87	
stroke_ever	0	1	0.06	0.23	0.00	0.00	0.00	0.00	1.00	
social_origins	0	1	0.79	1.10	0.00	0.00	0.00	1.00	4.00	
ses	0	1	0.37	0.63	-6.01	0.00	0.38	0.78	2.99	

Blah

```
#construct our analytic sample (via listwise deletion)
```