A4 / LEARN Data Primer

Introduction

This document demonstrates an example on how to explore the various clinical datasets made available using R programming language.

Load required R packages

```
library(tidyverse)
library(arsenal)
```

Load CSV data files into R

```
datadic <- read.csv("clinical_datadic.csv")     # Data dictionary
# raw dataset
ptdemog <- read.csv("ptdemog.csv")     # Participant demographics
# derived datasets
subjinfo <- read.csv("SUBJINFO.csv") # Subject info (re-screens and APOE status)
pacc <- read.csv("PACC.csv") # PACC
# imaging datasets
petsuvr <- read.csv("imaging_SUVR_amyloid.csv") # Amyloid PET Quantitative results
petvadata <- read.csv("imaging_PET_VA.csv") # Amyloid PET Eligibility
vmri <- read.csv("imaging_volumetric_mri.csv") # Volumetric MRI
```

Define functions for extracting meta data from data dictionary

```
get_levels <- function(x){
  as.numeric(unlist(lapply(strsplit(unlist(strsplit(subset(
    datadic, FIELD_NAME==x)$FIELD_CODE, ';')), '='), function(y) y[1])))
}
get_labels <- function(x){
  unlist(lapply(strsplit(unlist(strsplit(subset(
    datadic, FIELD_NAME==x)$FIELD_CODE, ';')), '='), function(y) y[2]))
}
```

Prepare data

Rescreens

Participants who re-screened may appear in the data twice with different BIDs each time. The SUBJINFO derived dataset indicates which participants are re-screens, and how the re-screen BIDs are mapped to each other. The code below also accounts for this to set up the removal of duplicate appearances.

```
rescreens <- subjinfo %>%
  filter(!is.na(PREVBID)) %>%
  rename(BID1 = PREVBID, BID2 = BID) %>%
  select(BID1, BID2)
```

Gather Amyloid PET data

The Amyloid PET quantitative data (`petsuvr`) is in long format with one row per region. We use `tidyr::pivot_wider` to transform to wide format with one column per region.

```
pet <- petsuvr %>%
  filter(brain_region != '' & VISCODE == 2) %>%
  pivot_wider(id_cols='BID', names_from=brain_region, values_from=suvr_cer) %>%
  left_join(petvadata, by='BID') %>%
  left_join(rescreens, by=c('BID' = 'BID1')) %>%
  mutate( # update PET BIDs to second BID if necessary
    BID = case_when(
      !is.na(BID2) ~ BID2,
```

```
      TRUE ~ BID)) %>%
  arrange(BID, BID2) %>%
  filter(!duplicated(BID, fromLast = TRUE)) %>%
  select(-BID2)
```

Label Participant Demographics data

Raw datasets are presented using coded values. Their translated labels can be found using the relevant data dictionary. The code below demonstrates how the labelled values can be derived using the functions previously defined at the start of this document.

```
ptdemog <- ptdemog %>%
  mutate(
    PTGENDER = factor(PTGENDER,
      levels = get_levels('PTGENDER'),
      labels = get_labels('PTGENDER')),
    PTETHNIC = factor(PTETHNIC,
      levels = get_levels('PTETHNIC'),
      labels = get_labels('PTETHNIC')),
    PTMARRY = factor(PTMARRY,
      levels = get_levels('PTMARRY'),
      labels = get_labels('PTMARRY')),
    PTNOTRT = factor(PTNOTRT,
      levels = get_levels('PTNOTRT'),
      labels = get_labels('PTNOTRT')))
```

Some variables were collected as multi-checkbox selections. The data is aggregated to a string of numeric values seperated by a colon (e.g. '2:3:5'). The code below is an example of how to convert the variable to distinct binomial fields.

```
race.levs <- get_levels('PTRACE')
race.labs <- get_labels('PTRACE')
for (i in 1:length(race.levs)){
  ptdemog[!is.na(ptdemog$PTRACE),  paste('PTRACE:',race.labs[i])] <- 0
  ptdemog[grep(race.levs[i], ptdemog$PTRACE, fixed = TRUE), paste('PTRACE:',race.labs[i])] <- 1
}
```

NOTE: Key participant demographics and baseline characteristics can also be found in the SUBJINFO derived dataset. PTDEMOG was used here to demonstrate how to label raw data using the data dictionary.

Baseline PACC

```
pacc_bl <- pacc %>%
  filter(VISCODE == 6) %>%
  select(BID, PACC.raw, MMSCORE, LDELTOTAL, DIGITTOTAL, FCTOTAL96)
```

Prepare data table

```
dd <- subjinfo[,c('BID','APOEGN')] %>%
  left_join(ptdemog, by='BID') %>%
  left_join(pet, by=c('SUBSTUDY','BID')) %>%
  left_join(vmri[is.na(vmri$VISCODE) | vmri$VISCODE==4,]) %>%
  left_join(pacc_bl, by=c('BID')) %>%
  filter(!BID %in% rescreens$BID1) %>% # remove first appearance of rescreens
  rename(`A4 amyloid eligibility` = overall_score) %>%
  rename(
    `Age at screening (yrs)` = PTAGE,
    `APOE genotype` = APOEGN,
    `Education (yrs)` = PTEDUCAT,
    `PET SUVr` = Composite_Summary,
    Sex = PTGENDER,
    Ethnicity = PTETHNIC,
    `Marital status` = PTMARRY,
    `Participant retired` = PTNOTRT,
    `Hippocampal Occupancy` = HOC,
    PACC = PACC.raw,
    MMSE = MMSCORE,
    `Logical memory delay` = LDELTOTAL,
    `Digit symbol` = DIGITTOTAL,
    `FCSRT (2xFree + Cued)` = FCTOTAL96
  ) %>%
  filter(!SUBSTUDY %in% 'SF') # excluding screen-fails
```

Summarize data

```
dd<- dd %>% # rename values and fields for footnote setup in summary table
  mutate(SUBSTUDY = ifelse(SUBSTUDY == 'A4', 'A4^a^' , SUBSTUDY)) %>%
  rename(
    `A4 amyloid eligibility^b^` = `A4 amyloid eligibility`,
    `Amyloid PET SUVr^b^` = `PET SUVr`
  )
tableby(SUBSTUDY ~ .,
  data = select(dd, SUBSTUDY, `A4 amyloid eligibility^b^`, `Age at screening (yrs)`, `Education (yrs)`,
    `APOE genotype`, `Amyloid PET SUVr^b^`, `Hippocampal Occupancy`, Sex, Ethnicity,
    `Marital status`, `Participant retired`,
    PACC, MMSE, `Logical memory delay`, `Digit symbol`, `FCSRT (2xFree + Cued)`), digits = 2) %>%
summary(title = "Baseline characteristics of A4-randomized^a^ and LEARN-enrolled cohorts")
```

Table 1: Baseline characteristics of A4-randomized[a] and LEARN-enrolled cohorts

	A4[a] (N=1169)	LEARN (N=539)	Total (N=1708)	p value
A4 amyloid eligibility[b]				< 0.001
negative	0 (0.0%)	539 (100.0%)	539 (31.6%)	
positive	1169 (100.0%)	0 (0.0%)	1169 (68.4%)	
Age at screening (yrs)				< 0.001
Mean (SD)	71.92 (4.81)	70.53 (4.32)	71.48 (4.70)	
Range	65.00 - 85.74	65.00 - 85.60	65.00 - 85.74	
Education (yrs)				0.123
Mean (SD)	16.57 (2.81)	16.79 (2.63)	16.64 (2.75)	
Range	7.00 - 30.00	8.00 - 30.00	7.00 - 30.00	
APOE genotype				< 0.001
N-Miss	0	2	2	
E2/E2	2 (0.2%)	5 (0.9%)	7 (0.4%)	
E2/E3	61 (5.2%)	67 (12.5%)	128 (7.5%)	
E2/E4	35 (3.0%)	10 (1.9%)	45 (2.6%)	
E3/E3	417 (35.7%)	342 (63.7%)	759 (44.5%)	
E3/E4	560 (47.9%)	111 (20.7%)	671 (39.3%)	
E4/E4	94 (8.0%)	2 (0.4%)	96 (5.6%)	
Amyloid PET SUVr[b]				< 0.001
Mean (SD)	1.33 (0.18)	0.99 (0.07)	1.22 (0.22)	
Range	0.97 - 2.09	0.79 - 1.16	0.79 - 2.09	
Hippocampal Occupancy				0.012
N-Miss	2	3	5	
Mean (SD)	0.70 (0.40)	0.75 (0.08)	0.72 (0.34)	
Range	-4.00 - 0.90	0.40 - 0.88	-4.00 - 0.90	
Sex				0.467
Male	475 (40.6%)	209 (38.8%)	684 (40.0%)	
Female	694 (59.4%)	330 (61.2%)	1024 (60.0%)	
Ethnicity				0.821
Hispanic or Latino	34 (2.9%)	18 (3.3%)	52 (3.0%)	
Not Hispanic or Latino	1124 (96.2%)	517 (95.9%)	1641 (96.1%)	
Unknown or Not reported	11 (0.9%)	4 (0.7%)	15 (0.9%)	
Marital status				0.155
Married	836 (71.5%)	386 (71.6%)	1222 (71.5%)	
Widowed	102 (8.7%)	53 (9.8%)	155 (9.1%)	
Divorced	170 (14.5%)	67 (12.4%)	237 (13.9%)	
Never married	42 (3.6%)	29 (5.4%)	71 (4.2%)	
Unknown/Other	19 (1.6%)	4 (0.7%)	23 (1.3%)	
Participant retired				0.693
Yes	877 (75.0%)	412 (76.4%)	1289 (75.5%)	
No	274 (23.4%)	121 (22.4%)	395 (23.1%)	
Not Applicable	18 (1.5%)	6 (1.1%)	24 (1.4%)	
PACC				< 0.001
Mean (SD)	-0.00 (2.68)	0.79 (2.35)	0.25 (2.60)	
Range	-12.52 - 7.75	-8.70 - 6.64	-12.52 - 7.75	
MMSE				< 0.001
Mean (SD)	28.78 (1.28)	29.03 (1.17)	28.86 (1.25)	
Range	22.00 - 30.00	23.00 - 30.00	22.00 - 30.00	
Logical memory delay				< 0.001
Mean (SD)	12.62 (3.68)	13.54 (3.35)	12.91 (3.61)	
Range	0.00 - 23.00	3.00 - 24.00	0.00 - 24.00	
Digit symbol				0.012
Mean (SD)	48.64 (10.02)	49.95 (9.89)	49.05 (10.00)	
Range	15.00 - 86.00	0.00 - 79.00	0.00 - 86.00	
FCSRT (2xFree + Cued)				< 0.001
Mean (SD)	77.35 (6.29)	78.66 (5.83)	77.76 (6.18)	
Range	44.00 - 92.00	58.00 - 94.00	44.00 - 94.00	

[a] A4-randomized cohort (n=1169) includes other participants in addition to the modified intention-to-treat population (mITT n=1147) reported in the A4 trial (Sperling et al. 2023). The modified intention-to-treat population population that was reported for the A4 trial results include those who received at least one dose of solanezumab or placebo and underwent assessment for the primary end point. Please refer to the Intro-to-A4.pdf file for code to reproduce the baseline characteristics and primary findings of the A4 study.

[b] Refer to A4 Amyloid PET Eligibility Methods PDF document (imaging_PET_VA_methods.pdf) regarding these amyloid-related measures, the eligibility determination process and the modifications made to the SUVR algorithm.

References

R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Sperling, R. A., Donohue, M. C., Raman, R., Rafii, M. S., Johnson, K., Masters, C. L., van Dyck, C. H., Iwatsubo, T., Marshall, G. A., Yaari, R., Mancini, M., Holdridge, K. C., Case, M., Sims, J. R., Aisen, P. S., & A4 Study Team (2023). Trial of Solanezumab in Preclinical Alzheimer's Disease. *The New England journal of medicine, 389*(12), 1096–1107. https://doi.org/10.1056/NEJMoa2305032