01_table1_demographic

randy

2023-10-24

- Table1 has three different versions :
  - the basic training & testing
  - the basic training & testing + overall
  - the basic training & testing + pvalues
- Do we need to include too many variables not used in the paper?
  - ethnicity
  - genotype

Setup

```r
## set seed
set.seed(555)
# load("data/sysdata.rda")
# load("final/epic_clean_full_data.Rdata")

# the code to prepare for
# epic, demog, test and train
# data0, data1, and data2
# they are all saved in sysdata.rda files
data <- left_join(epic, demog, by = "id") %>%
  mutate(sex = as.factor(sex))
test_id <- unique(test$id) %>% unlist()
train_id <- unique(train$id) %>% unlist()

data0 <- data %>%
  mutate(group = case_when(id %in% test_id ~ "testing",
                           TRUE ~ "training"))
data1 <- data0 %>%
  group_by(id, group) %>%
  summarize(age_mean = mean(age),
            age_min = min(age),
            age_max = max(age),
            age_n = length(age),
            visitn = n(),
            h_mean = mean(ht),
            h_max = max(ht),
            h_min = min(ht),
            w_mean = mean(wt),
            w_max = max(wt),
            w_min = min(wt),
            sex = sex,
            genotype = genotype,
            ethnic = ethnic,
            race = race) %>%
  ungroup() %>%
  unique()

# working dataset
data2 <- full_join(data1, data,
                   by = join_by(id, sex, genotype, ethnic, race)) %>%
  as.data.frame() %>%
  mutate(time = age - age_min,
         age_diff = age_max - age_min)

write.csv(data1, file = paste0("data/S01_table1_dataset_randy_", Sys.Date(), ".csv"))
write.csv(data2, file = paste0("data/S01_epic_clean_randy_", Sys.Date(), ".csv"))
```

Making table1

```r
library(readr)
data1 <- read_csv("data/S01_table1_dataset_randy_2023-08-23.csv")


## table0 contains all the information about demgo for total
table0 <- data1 %>%
  unique() %>%
  dplyr::select(-id) %>%
  mutate(ethnic = case_when(ethnic == 1 ~ "Hispanic",
                            ethnic == 2 ~ "Non-Hispanic"),
         race = case_when(race == 1 ~ "White",
                          race != 1 ~ "Other"),
         sex = case_when(sex == "F" ~ "Female",
                         sex == "M" ~ "Male"),
         age_diff = age_max - age_min) %>%
  dplyr::select(group,
                Genotype = genotype,
                Gender = sex,
                Race = race,
                Ethnicity = ethnic,
                "Visit number" = visitn,
                # "Age mean" = age_mean,
                "Age baseline" = age_min,
                "Age final" = age_max,
                "Follow up years" = age_diff,
                # "Height mean" = h_mean,
                "Height baseline" = h_min) %>%
  # "Weight mean" = w_mean,
  # "Weight baseline" = w_min
  ## select all the variables for table1
  tbl_summary(by = group) %>%
  ## just display all the variables in one column
  modify_header(label = "**Characteristics**") %>%
  # update the column header
  bold_labels() %>%
  italicize_labels() %>%
  # as.data.frame()
  as_flex_table() %>%
  flextable::bold(part = "header") %>%
  ## auto adjust the column widths
  flextable::autofit()

## table1 contains information of dataset grouped as training and testing
table1 <- data1 %>%
  unique() %>%
  dplyr::select(-id) %>%
  mutate(ethnic = case_when(ethnic == 1 ~ "Hispanic",
                            ethnic == 2 ~ "Non-Hispanic"),
         race = case_when(race == 1 ~ "White",
                          race != 1 ~ "Other"),
         sex = case_when(sex == "F" ~ "Female",
                         sex == "M" ~ "Male"),
         age_diff = age_max - age_min) %>%
  dplyr::select(group,
                Genotype = genotype,
                Gender = sex,
                Race = race,
                Ethnicity = ethnic,
                "Visit number" = visitn,
                # "Age mean" = age_mean,
                "Age baseline" = age_min,
                "Age final" = age_max,
                "Follow up years" = age_diff,
                # "Height mean" = h_mean,
                "Height baseline" = h_min) %>%
  # "Weight mean" = w_mean,
  # "Weight baseline" = w_min)
  ## select all the variables for table1
  tbl_summary(by = group) %>%
  ## just display all the variables in one column
  modify_header(label = "**Characteristics**") %>%
  # update the column header
  bold_labels() %>%
  add_p() %>%
  italicize_labels() %>%
  as.data.frame()


## table1 contains information of dataset grouped as training and testing
table2 <- data1 %>%
  unique() %>%
  dplyr::select(-id) %>%
  mutate(ethnic = case_when(ethnic == 1 ~ "Hispanic",
                            ethnic == 2 ~ "Non-Hispanic"),
         race = case_when(race == 1 ~ "White",
                          race != 1 ~ "Other"),
         sex = case_when(sex == "F" ~ "Female",
                         sex == "M" ~ "Male"),
         age_diff = age_max - age_min) %>%
  dplyr::select(group,
                Genotype = genotype,
                Gender = sex,
                Race = race,
                Ethnicity = ethnic,
                "Visit number" = visitn,
                # "Age mean" = age_mean,
                "Age baseline" = age_min,
                # "Age final" = age_max,
                "Follow up years" = age_diff,
                # "Height mean" = h_mean,
                "Height baseline" = h_min) %>%
  # "Weight mean" = w_mean,
  # "Weight baseline" = w_min)
  ## select all the variables for table1
  tbl_summary(by = group,
              statistic = list(all_continuous() ~ "{mean} ({sd})") ) %>%
  ## just display all the variables in one column
  modify_header(label = "**Characteristics**") %>%
  # update the column header
  bold_labels() %>%
  add_overall(last = TRUE) %>%
  italicize_labels()
```

Saving for the table1

```
sessionInfo()
```

```
#> R version 4.2.2 (2022-10-31)
#> Platform: aarch64-apple-darwin20 (64-bit)
#> Running under: macOS 14.0
#>
#> Matrix products: default
#> BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib
#>
#> locale:
#> [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#>  [1] flextable_0.9.2 gtsummary_1.7.1 lubridate_1.9.2 forcats_1.0.0
#>  [5] stringr_1.5.0   dplyr_1.1.2     purrr_1.0.1     readr_2.1.4
#>  [9] tidyr_1.3.0     tibble_3.2.1    ggplot2_3.4.3   tidyverse_2.0.0
#> [13] here_1.0.1
#> 
#> loaded via a namespace (and not attached):
#>  [1] Rcpp_1.0.11             freshr_1.0.2            rprojroot_2.0.3
#>  [4] digest_0.6.33           utf8_1.2.3              mime_0.12
#>  [7] R6_2.5.1                backports_1.4.1         evaluate_0.21
#> [10] pillar_1.9.0            gdtools_0.3.3           rlang_1.1.1
#> [13] uuid_1.1-0              curl_5.0.1              rstudioapi_0.15.0
#> [16] data.table_1.14.8       rmarkdown_2.23          textshaping_0.3.6
#> [19] bit_4.0.5               munsell_0.5.0           broom_1.0.5
#> [22] shiny_1.7.4.1           compiler_4.2.2          httpuv_1.6.11
#> [25] xfun_0.39               askpass_1.1             pkgconfig_2.0.3
#> [28] systemfonts_1.0.4       gfonts_0.2.0            htmltools_0.5.5
#> [31] openssl_2.1.0           tidyselect_1.2.0        fontBitstreamVera_0.1.1
#> [34] httpcode_0.3.0          fansi_1.0.4             crayon_1.5.2
#> [37] tzdb_0.4.0              withr_2.5.0             later_1.3.1
#> [40] crul_1.4.0              grid_4.2.2              jsonlite_1.8.7
#> [43] xtable_1.8-4            gtable_0.3.3            lifecycle_1.0.3
#> [46] magrittr_2.0.3          scales_1.2.1            zip_2.3.0
#> [49] vroom_1.6.3             cli_3.6.1               stringi_1.7.12
#> [52] broom.helpers_1.13.0    promises_1.2.0.1        xml2_1.3.5
#> [55] ragg_1.2.5              ellipsis_0.3.2          generics_0.1.3
#> [58] vctrs_0.6.3             tools_4.2.2             bit64_4.0.5
#> [61] glue_1.6.2              officer_0.6.2           fontquiver_0.2.1
#> [64] hms_1.1.3               parallel_4.2.2          fastmap_1.1.1
#> [67] yaml_2.3.7              timechange_0.2.0        colorspace_2.1-0
#> [70] fontLiberation_0.1.0    gt_0.9.0                knitr_1.43
```