User Guide

Liam Wright

2023-07-15

This user guides describes the `ncds.Rds` dataset available in this ReShare repository. The dataset was created for the paper, '*Many Models in R: A Tutorial*'. `ncds.Rds` is an `R` format synthetic dataset created with the `synthpop` dataset in `R` using data from the National Child Development Study (NCDS), a birth cohort of individuals born in a single week of March 1958 in Britain (Power and Elliott 2006; University College London, n.d.). The data is only intended to be used in the tutorial - it is not to be used for drawing statistical inferences.

The dataset contains data on fourteen biomarkers collected at the age 46/47 sweep of the survey, four measures of cognitive ability from age 11 and 16, and three covariates, sex, body mass index at age 11 and father's social class. The 14 biomarkers are used in Castagné et al. (2018) and the measures of cognitive ability are used in Bann et al. (2023) . A data dictionary is provided in Appendix 1.

All variables have received rudimentary cleaning, with missing values imputed using the chained equations method with the `mice` package in R. The biomarkers and cognitive ability variables have each been standardized (mean = 0, standard deviation = 1) for comparability. HDL cholesterol, salivary cortisol change, insulin-like growth factor-1, and peak expiratory flow have also been reversed so that higher values of all biomarkers represent greater risk of ill health. The code used to clean the raw data, impute missing values, and generate a synthetic dataset is presented in Appendix 2. To repeat, the data are not intended to be used in proper statistical analyses. Instead, they are for tutorial purposes only.

References

Bann, David, Liam Wright, Neil M Davies, and Vanessa Moulton. 2023. "Weakening of the Cognition and Height Association from 1957 to 2018: Findings from Four British Birth Cohort Studies." *eLife* 12 (April): e81099. https://doi.org/10.7554/eLife.81099.

Castagné, Raphaële, Valérie Garès, Maryam Karimi, Marc Chadeau-Hyam, Paolo Vineis, Cyrille Delpierre, Michelle Kelly-Irving, and Lifepath Consortium. 2018. "Allostatic Load and Subsequent All-Cause Mortality: Which Biological Markers Drive the Relationship? Findings from a UK Birth Cohort." *European Journal of Epidemiology* 33 (5): 441–58. https://doi.org/10.1007/s10654-018-0364-1.

Power, Chris, and Jane Elliott. 2006. "Cohort Profile: 1958 British Birth Cohort (National Child Development Study)." *International Journal of Epidemiology* 35 (1): 34–41. https://doi.org/10.1093/ije/dyi183.

University College London. n.d. "National Child Development Study." https://doi.org/10.5255/UKDA-SERIES-2000032.

Appendix 1: Data Dictionary

pos	variable	label	col_type	levels
1	id	ID (Synthetic)	int	
2	bio_cortisol_baseline	Salivary Cortisol (Baseline Level)	dbl	

pos	variable	label	col_type	levels
3	bio_cortisol_difference	Salivary Cortisol (Change)	dbl	
4	bio_igf_1	Insulin-Like Growth Factor-1 (IGF-1)	dbl	
5	bio_crp	C-Reactive Protein (CRP)	dbl	
6	bio_fibrinogen	Fibrinogen	dbl	
7	bio_ige	Immunoglobulin E (IgE)	dbl	
8	bio_hdl	High-Density Lipoprotein (HLD) Cholesterol	dbl	
9	bio_ldl	Low-density Lipoprotein (LDL) Cholesterol	dbl	
10	bio_trig	Triglycerides	dbl	
11	bio_hba1c	Glycosylated Haemoglobin (HbA1C)	dbl	
12	bio_sbp	Systolic Blood Pressure (SBP)	dbl	
13	bio_dbp	Diastolic Blood Pressure (DBP)	dbl	
14	bio_heart_rate	Heart Rate	dbl	
15	bio_peak_flow	Peak Expiratory Flow	dbl	
16	cog_maths_11	Maths Score @ Age 11	dbl	
17	cog_maths_16	Maths Score @ Age 16	dbl	
18	cog_verbal_11	Verbal Score @ Age 16	dbl	
19	cog_vocab_16	Vocabulary	dbl	
20	sex	Sex (Ref: Male)	fct	Male, Female
21	bmi_11	Body Mass Index @ Age 11	dbl	
22	father_class	Father's Social Class @ Age 11	fct	I, II, III non manual, III manual, IV, V

Appendix 2: Code to Generate the Synthetic Data

```
library(tidyverse)
library(haven)
library(labelled)
library(glue)
library(synthpop)
library(summarytools)
library(dlookr)
library(mice)
library(knitr)

rm(list = ls())

# 1. Data Exploration Functions ----
ncds_fld <- Sys.getenv("ncds_fld")  # File path to where raw NCDS files are

negative_to_na <- function(x){
  na_range(x) <- c(-Inf, -1)
  user_na_to_na(x)
}
neg <- function(x){
  -as.double(x)
}
```

```
# 2. Load Data ----
## a. Biomarkers ----
bio_raw <- glue("{ncds_fld}/42y-44y Biomedical/ncds42-4_biomedical_eul.dta") %>%
  read_dta()

bio_clean <- bio_raw %>% #select(fvc:max1) %>% descr() %>% tb() %>% View()
  transmute(
    id = ncdsid,

    # Neuroendocrine system
    across(c(cortred, cortblue), negative_to_na),
    cortisol_baseline = cortred,
    cortisol_difference = neg(cortred - cortblue),

    # Immune and inflammatory system
    igf_1 = negative_to_na(igf1) %>% neg(),
    crp = ifelse(between(crp, 0, 10), crp, NA),
    fibrinogen = negative_to_na(fib),
    ige = negative_to_na(ige),

    # Metabolic system
    hdl = negative_to_na(hdl) %>% neg(),
    ldl = negative_to_na(ldl),
    trig = negative_to_na(trig),
    hba1c = negative_to_na(hba1c),

    # Cardiovascular system
    across(c(sys, sys2, sys3), ~ ifelse(between(.x, 0, 998), .x, NA)),
    sbp = (sys + sys2 + sys3) / 3,
    across(c(dias, dias2, dias3), ~ ifelse(between(.x, 0, 998), .x, NA)),
    dbp = (dias + dias2 + dias3) / 3,
    across(c(pulse, pulse2, pulse3), ~ ifelse(between(.x, 0, 998), .x, NA)),
    heart_rate = (pulse + pulse2 + pulse3) / 3,
    peak_flow = negative_to_na(htpf) %>% neg()

  ) %>%
  select(id,
         cortisol_baseline, cortisol_difference,
         igf_1, crp, fibrinogen, ige,
         hdl, ldl, trig, hba1c,
         sbp, dbp, heart_rate, peak_flow) %>%
  rename_with(~ glue("bio_{.x}"), .cols = -id)

descr(bio_clean) %>% tb()

## b. Cognitive Ability ----
cog_raw <- glue("{ncds_fld}/0y-16y/ncds0123.dta") %>%
  read_dta()

cog_clean <- cog_raw %>%
  transmute(
    id = ncdsid,
```

```
    sex = negative_to_na(n622) %>% as_factor(),

    bmi_11 = negative_to_na(dvwt11)/(negative_to_na(dvht11)^2),

    cog_maths_11 = negative_to_na(n926),
    cog_maths_16 = negative_to_na(n2930),
    cog_verbal_11 = negative_to_na(n914),
    cog_vocab_16 = negative_to_na(n2928),

    father_class = negative_to_na(n1687) %>%
      as_factor() %>%
      fct_recode(NULL = "No male head")

  )

# 3. Impute Data ----
df_miss <- bio_clean %>%
  left_join(cog_clean, by = "id") %>%
  zap_label() %>%
  zap_labels() %>%
  zap_formats()

df_imp <- mice(df_miss %>% select(-id),
               m = 1,
               method = "rf",
               seed = 1) %>%
  complete()

scale <- function(x){
  (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)
}

set.seed(1)
df_syn <- syn(df_imp)

replicated.uniques(df_syn, df_imp)$no.replications

df <- df_syn %>%
  pluck("syn") %>%
  as_tibble() %>%
  mutate(id = row_number(),
         .before = 1) %>%
  group_by(sex) %>%
  mutate(across(matches("^bio_"), scale)) %>%
  ungroup() %>%
  mutate(across(matches("^cog_"), scale)) %>%
  relocate(sex, bmi_11, .after = cog_vocab_16) %>%
  set_variable_labels(
    id = "ID (Synthetic)",
    bio_cortisol_baseline = "Salivary Cortisol (Baseline Level)",
    bio_cortisol_difference = "Salivary Cortisol (Change)",
    bio_igf_1  = "Insulin-Like Growth Factor-1 (IGF-1)",
    bio_crp  = "C-Reactive Protein (CRP)",
    bio_fibrinogen  = "Fibrinogen",
```

```
    bio_ige  = "Immunoglobulin E (IgE)",
    bio_hdl  = "High-Density Lipoprotein (HLD) Cholesterol",
    bio_ldl  = "Low-density Lipoprotein (LDL) Cholesterol",
    bio_trig  = "Triglycerides",
    bio_hba1c  = "Glycosylated Haemoglobin (HbA1C)",
    bio_sbp  = "Systolic Blood Pressure (SBP)",
    bio_dbp  = "Diastolic Blood Pressure (DBP)",
    bio_heart_rate  = "Heart Rate",
    bio_peak_flow  = "Peak Expiratory Flow",
    cog_maths_11  = "Maths Score @ Age 11",
    cog_maths_16  = "Maths Score @ Age 16",
    cog_verbal_11  = "Verbal Score @ Age 16",
    cog_vocab_16  = "Vocabulary",
    sex  = "Sex (Ref: Male)",
    bmi_11  = "Body Mass Index @ Age 11",
    father_class  = "Father's Social Class @ Age 11",
  )

saveRDS(df, file = "Code/ncds.Rds")

descr(df) %>% tb()
```