Data cleaning guide for the Rotterdam Study

Paloma Rojas Saunero, Eline Vinke

2019-07-23

Contents

Chapter 1

Prerequisites

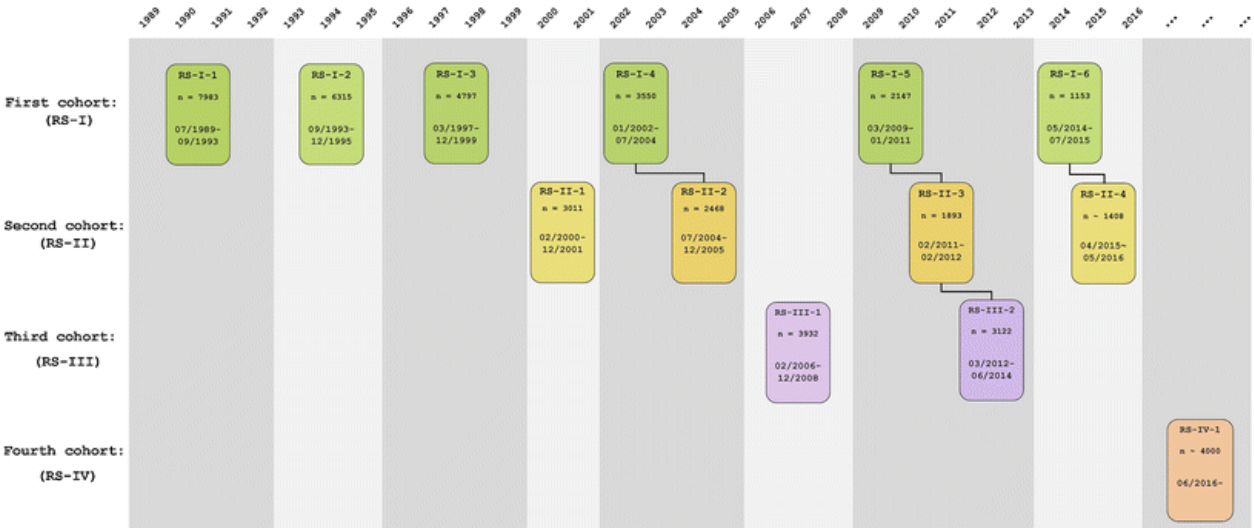
- Install R and Rstudio IDE
- Install and open the following packages:
- Create an R project with a folder for your raw data (We will name this folder `00_raw_data`)
- Data can be accessed from the following: link

Chapter 2

Introduction

We will use the *tidyverse* language and ecosystem. The intentation is that all students becoming familiar with the data, develop the essencial skills for a reproducible data cleaning, since this is the first step before doing any data analysis.

We followed the structure of the visit process to name the cohorts and visits:



In the following structure:

var1	var2	var3	var4	var5	var6
RS-I-1	RS-I-2	RS-I-3	RS-I-4	RS-I-5	RS-I-6
NA	NA	RS-II-1 (ep)	RS-II-2	RS-III-3	RS-III-4
NA	NA	NA	RS-III-1 (ej)	RS-III-2	RS-III-3

Each small chapter will provide the code to clean each of the variables that are oftenly used as covariates in any data analysis.

Since the name of the variables could have changed in each visit, and data is arranged in different ways, we decided to create a systematic flow of how the cleaning process is one covariate at a time, for all visits, for all cohorts.

The usuals steps will include:

1. Import datasets for all the visits, for all cohorts.

- Note that we use the `here` package to specify the folder/subfolder and file we want to import. We do this to avoid specifying the directory. This practice helps the reproducibility of the code as mentioned in the following link
  - We will always include the subfix representing the cohort and visit. For example:

```
rs1_1 <- read_sav(here::here("00_raw_data", "visits", "Ergo1ResponseDetail_(22-jan-2015)_excerpt.sav"))
rs1_2 <- read_sav(here::here("00_raw_data", "visits", "Ergo2ResponseDetail_(22-jan-2015)_excerpt.sav"))
rs1_3 <- read_sav(here::here("00_raw_data", "visits", "e3_(3)_RESPONS_(22-feb-2016)_excerpt.sav"))
rs1_4 <- read_sav(here::here("00_raw_data", "visits", "e4_(4)_RESPONS_(12-mar-2018)_excerpt.sav"))
rs1_5 <- read_sav(here::here("00_raw_data", "visits", "e5_(5)_RESPONS_(22-jun-2016)_excerpt.sav"))
rs1_6 <- read_sav(here::here("00_raw_data", "visits", "e6_(6)_RESPONS_(10-feb-2017)_EXCERPT.sav"))
rs2_1 <- read_sav(here::here("00_raw_data", "visits", "ep_(1)_RESPONS_(15-jan-2019)_excerpt.sav"))
rs3_1 <- read_sav(here::here("00_raw_data", "visits", "ej_(1)_RESPONS_(04-apr-2016)_excerpt.sav"))
```

2. We will split the datasets that have data for more than one cohort, and name them by their respective cohort - visit. For example:

```
# Separate rs1_4 into rs1, rs2

rs1_4 <- rs1_4 %>%
  filter(rs_cohort == 1)

rs2_2 <- rs1_4 %>%
  filter(rs_cohort == 2)
```

3. Merge the data for all visits, by cohort:

```
rs1 <- list(rs1_1, rs1_2, rs1_3, rs1_4, rs1_5, rs1_6)

rs1_vis <- reduce(rs1, left_join, by = c("ergoid", "rs_cohort"))
```

4. Select the specific variables from the combined dataset, by cohort and rename for easier comprehension, by cohort:

```
rs1_bmi <- rs1_bmi%>%
  select(ergoid, rs_cohort, e1_aahgt, e1_aawgt, e2_229, e2_230, e3_229, e3_230, e4_229, e4_230, e5_229,
  rename(hgt1 = e1_aahgt, hgt2 = e2_229, hgt3 = e3_229, hgt4 = e4_229, hgt5 = e5_229, hgt6 = e6_229,
         wgt1 = e1_aawgt, wgt2 = e2_230, wgt3 = e3_230, wgt4 = e4_230, wgt5 = e5_230, wgt6 = e6_230)
```

5. Bind, if necessary, the cohorts. Since the variable names are consistent through the datasets.

```
rs_bmi <- rs1_bmi %>%
  bind_rows(rs2_bmi) %>%
  bind_rows(rs3_bmi)
```

6. Create new variables (Example)

```
rs_bmi <- rs_bmi %>%
  mutate(bmi1 = wgt1/((hgt1/100)^2),
         bmi2 = wgt2/((hgt2/100)^2),
         bmi3 = wgt3/((hgt3/100)^2),
         bmi4 = wgt4/((hgt4/100)^2),
         bmi5 = wgt5/((hgt5/100)^2),
         bmi6 = wgt6/((hgt6/100)^2))

#Note that we could have created a function, but the intention of this code is to make adaptable.
```

7. Export the variable to a `clean_data` folder.

```
export(bmi, here::here("02_clean_data", "bmi.Rdata"))
```

8. Merge variables by `ergoid` for a complete folder

Chapter 3

ERGO Basics

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Ergobasics

This data only contains baseline information for the three cohorts, so we can skip steps 2,3 and 6 described in 2

1. Import data

```
basic <- read_sav(here::here("00_raw_data", "basic", "RoterdamStudy_Basics2014.sav"))
```

2. Select variables:

```
basic <- basic %>%
  select(ergoid, rs_cohort, sex, date_of_birth, startdat)
```

3. Transform variables:

```
basic <- basic %>%
  mutate(age_0 = round(as.numeric(as.period((date_of_birth %--% startdat ), "years"), "years"), 2),
         sex = labelled(sex, c(Female = 1, Male = 0)))
```

4. Export:

```
export(basic, here::here("02_clean_data", "basic.Rdata"))
```

Chapter 4

ERGO Vital status

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Fp_mortality

This data only contains baseline information for the three cohorts, so we can skip steps 2,3 and 6 described in 2

1. Import data

```
vital_status <- import(here::here("00_raw_data", "vital_status", "fp_VitalStatus_(24-MAY-2018).sav"))
```

2. Select and rename variables:

```
vital_status <- vital_status %>%
  select(ergoid, fp_mortdat, fp_censordate) %>%
  rename(mort_date = fp_mortdat, censor_date = fp_censordate)
```

3. Export:

```
export(vital_status, here::here("02_clean_data", "vital_status.Rdata"))
```

Chapter 5

Visit dates

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Response_data

1. Import data for all cohorts

```
rs1_1 <- read_sav(here::here("00_raw_data", "visits", "Ergo1ResponseDetail_(22-jan-2015)_excerpt.sav"))
rs1_2 <- read_sav(here::here("00_raw_data", "visits", "Ergo2ResponseDetail_(22-jan-2015)_excerpt.sav"))
rs1_3 <- read_sav(here::here("00_raw_data", "visits", "e3_(3)_RESPONS_(22-feb-2016)_excerpt.sav"))
rs1_4 <- read_sav(here::here("00_raw_data", "visits", "e4_(4)_RESPONS_(12-mar-2018)_excerpt.sav"))
rs1_5 <- read_sav(here::here("00_raw_data", "visits", "e5_(5)_RESPONS_(22-jun-2016)_excerpt.sav"))
rs1_6 <- read_sav(here::here("00_raw_data", "visits", "e6_(6)_RESPONS_(10-feb-2017)_EXCERPT.sav"))
rs2_1 <- read_sav(here::here("00_raw_data", "visits", "ep_(1)_RESPONS_(15-jan-2019)_excerpt.sav"))
rs3_1 <- read_sav(here::here("00_raw_data", "visits", "ej_(1)_RESPONS_(04-apr-2016)_excerpt.sav"))
```

2. Split datasets:

```
#1. Separate rs1_4 into rs1, rs2

rs2_2 <- rs1_4 %>%
  filter(rs_cohort == 2)

rs1_4 <- rs1_4 %>%
  filter(rs_cohort == 1)

#2. Separate rs1_5 into rs1, rs2, rs3
rs3_2 <- rs1_5 %>%
  filter(rs_cohort == 3)

rs2_3 <- rs1_5 %>%
  filter(rs_cohort == 2)

rs1_5 <- rs1_5 %>%
  filter(rs_cohort == 1)

#3. rs1_6 into rs1 and rs2
rs2_4 <- rs1_6 %>%
  filter(rs_cohort == 2)

rs1_6 <- rs1_6 %>%
  filter(rs_cohort == 1)
```

3. Merge the data for all visits, by cohort:

```
### Merge RSI

rs1 <- list(rs1_1, rs1_2, rs1_3, rs1_4, rs1_5, rs1_6)

rs1_vis <- reduce(rs1, left_join, by = c("ergoid", "rs_cohort"))

### NOTE: RS1_2 had only 1 center visit and no home interview!

### Merge RS2
rs2 <- list(rs2_1, rs2_2, rs2_3, rs2_4)

rs2_vis <- reduce(rs2, left_join, by = c("ergoid", "rs_cohort"))

### Merge RS3

rs3_vis <- rs3_1  %>%
  left_join(rs3_2, by = c("ergoid", "rs_cohort"))
```

4. Select and rename the specific variables from the combined dataset, by cohort: In this case we selected the variables for the interview date.

```
rs1_interview <- rs1_vis %>%
  select(ergoid, rs_cohort, e1_aintdat, e2_bcendat, e3_3493, e4_3493, e5_3493, e6_3493) %>%
  rename(e1 = e1_aintdat, e2 = e2_bcendat, e3 = e3_3493, e4 = e4_3493, e5 = e5_3493, e6 = e6_3493)

rs2_interview <- rs2_vis %>%
  select(ergoid, rs_cohort , ep_3493, e4_3493, e5_3493, e6_3493) %>%
  rename(e3 = ep_3493, e4 = e4_3493, e5 = e5_3493, e6 = e6_3493)

rs3_interview <- rs3_vis %>%
  select(ergoid, rs_cohort , ej_3493, e5_3493) %>%
  rename(e4 = ej_3493, e5 = e5_3493)
```

5. Bind, if necessary, the cohorts. Since the variable names are consistent through the datasets.

```
visits <- rs1_interview %>%
  bind_rows(rs2_interview) %>%
  bind_rows(rs3_interview)
```

7. Export the variable to a `clean_data` folder.

```
export(visits, here::here("02_clean_data", "visits.Rdata"))
```

Chapter 6

BMI, weight and height

Source:

- https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/E1_anthropo
- https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/E2_anthropo
- https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/E5_anthropo

1. Import data for all cohorts

```
bmi1_1 <- read_sav(here::here("00_raw_data", "anthro", "/e1_ANTHROPO_(15-jun-2011).sav"))
bmi1_2 <- read_sav(here::here("00_raw_data", "anthro", "/e2_(2)_ANTHROPO_(26-apr-2011).sav"))
bmi1_3 <- read_sav(here::here("00_raw_data", "anthro", "/e3_(3)_HARTVAAT_(25-feb-2013)_ANTHROPO.sav"))
bmi1_4 <- read_sav(here::here("00_raw_data", "anthro", "/e4_(4)_UITSCHR_(06-nov-2014)_ANTHROPO-PART.sav
bmi1_5 <- read_sav(here::here("00_raw_data", "anthro", "/e5_(5)_ANTHROPO_(10-dec-2015).sav"))
bmi1_6 <- read_sav(here::here("00_raw_data", "anthro", "/e6_(6)_ANTHROPO_(25-apr-2017).sav"))
bmi2_1 <- read_sav(here::here("00_raw_data", "anthro", "/ep_(1)_LICHONDZ_(18-oct-2012)_ANTHROPO.sav"))
bmi3_1 <- read_sav(here::here("00_raw_data", "anthro", "/ej_(1)_UITSCHR_(23-feb-2010)_ANTHROPO-PART.sav
```

2. Split datasets:

```
# 1. Separate bmi1_4 into rs1 and rs2
bmi2_2 <- bmi1_4 %>%
  filter(rs_cohort == 2)

bmi1_4 <- bmi1_4 %>%
  filter(rs_cohort == 1)

# 2. Separate bmi1_5 into rs1, rs2, rs3
bmi3_2 <- bmi1_5 %>%
  filter(rs_cohort == 3)

bmi2_3 <- bmi1_5 %>%
  filter(rs_cohort == 2)

bmi1_5 <- bmi1_5 %>%
  filter(rs_cohort == 1)

# 3. Separate bmi1_6 into rs1 and rs2
bmi1_6 <- bmi1_6 %>%
  filter(rs_cohort ==1)
```

```
bmi2_4 <- bmi1_6 %>%
  filter(rs_cohort ==2)
```

3. Merge the data for all visits, by cohort:

```
# Merge cohorts for rs1

bmi1 <- list(bmi1_1, bmi1_2, bmi1_3, bmi1_4, bmi1_5, bmi1_6)

rs1_bmi <- reduce(bmi1, left_join, by = c("ergoid", "rs_cohort"))

# Merge cohorts for rs2

bmi2 <- list(bmi2_1, bmi2_2, bmi2_3, bmi2_4)

rs2_bmi <- reduce(bmi2, left_join, by = c("ergoid", "rs_cohort"))

# Merge cohorts for rs3

bmi3 <- bmi3_1 %>%
  left_join(bmi3_2, by= c("ergoid", "rs_cohort"))
```

4. Select and rename the specific variables from the combined dataset, by cohort: In this case we selected the variables for the interview date.

```
rs1_bmi <- rs1_bmi%>%
  select(ergoid, rs_cohort, e1_aahgt, e1_aawgt, e2_229,
         e2_230, e3_229, e3_230, e4_229, e4_230, e5_229,
         e5_230,e6_229,e6_230) %>%
  rename(hgt1 = e1_aahgt, hgt2 = e2_229, hgt3 = e3_229,
         hgt4 = e4_229, hgt5 = e5_229, hgt6 = e6_229,
         wgt1 = e1_aawgt, wgt2 = e2_230, wgt3 = e3_230,
         wgt4 = e4_230, wgt5 = e5_230, wgt6 = e6_230)

rs2_bmi <- rs2_bmi%>%
  select(ergoid, rs_cohort, ep_229, ep_230, e4_229,
         e4_230, e5_229, e5_230,e6_229,e6_230) %>%
  rename(hgt3 = ep_229, hgt4 = e4_229, hgt5 = e5_229, hgt6 = e6_229,
         wgt3 = ep_230, wgt4 = e4_230, wgt5 = e5_230, wgt6 = e6_230)


rs3_bmi <- bmi3%>%
  select(ergoid,  rs_cohort, ej_229, ej_230, e5_229, e5_230) %>%
  rename(hgt4 = ej_229, hgt5 = e5_229,
         wgt4 = ej_230, wgt5 = e5_230)
```

5. Bind, if necessary, the cohorts. Since the variable names are consistent through the datasets.

```
rs_bmi <- rs1_bmi %>%
  bind_rows(rs2_bmi) %>%
  bind_rows(rs3_bmi)
```

6. Create `bmi` variable

```
rs_bmi <- rs_bmi %>%
  mutate(bmi1 = wgt1/((hgt1/100)^2),
```

```
          bmi2 = wgt2/((hgt2/100)^2),
          bmi3 = wgt3/((hgt3/100)^2),
          bmi4 = wgt4/((hgt4/100)^2),
          bmi5 = wgt5/((hgt5/100)^2),
          bmi6 = wgt6/((hgt6/100)^2))
```

7. Export the variable to a `clean_data` folder.

```
export(rs_bmi, here::here("02_clean_data", "bmi.Rdata"))
```

Chapter 7

Education

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Education

This data only contains baseline information for the three cohorts, so we can skip steps 2,3 and 6 described in 2

1. Import data

```
educ <- read_sav(here::here("00_raw_data", "education", "Education RS-I-II-III (UNESCO class)_(12-MAR-2(
```

2. Rename variable:

```
educ <- educ %>%
  rename(education = ses_UNESCO_recoded)
```

3. Add labels to variable

```
educ <- educ %>%
  mutate(education = labelled(
    education,
    c(
      `Primary education` = 0,
      `lower or intermediate general education OR lower vocational education` = 1,
      `intermediate vocational education OR higher general education` = 2,
      `higher vocational education OR university` = 3)
  ))
```

4. Export:

```
export(educ, here::here("02_clean_data", "educ.Rdata"))
```

Chapter 8

Alcohol

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Alcohol

1. Import data for all cohorts

```
oh1_1 <- read_sav(here::here("00_raw_data", "alcohol", "e1_FFQ_AlcoholGramPerday_inclGLAZEN.sav"))
oh1_2 <- read_sav(here::here("00_raw_data", "alcohol", "e2_intvw_Alcoholperday_25-10-2013.sav"))
oh1_3 <- read_sav(here::here("00_raw_data", "alcohol", "e3_intvw_Alcoholperday_24-10-2017.sav"))
oh1_4 <- read_sav(here::here("00_raw_data", "alcohol", "e4_intvw_Alcoholperday_22-11-2013.sav"))
oh1_5 <- read_sav(here::here("00_raw_data", "alcohol", "e5intvw_Alcoholperday_11-07-2014.sav"))
oh1_6 <- read_sav(here::here("00_raw_data", "alcohol", "e6_AlcoholGramsDay_FFQ_energy_(13-dec-2018).sav
oh2_1 <- read_sav(here::here("00_raw_data", "alcohol", "ep_intvw_Alcoholperday_22-11-2013.sav"))
oh3_1 <- read_sav(here::here("00_raw_data", "alcohol", "ej_intvw_ALCOHOLGRAMSPERDAY_(14072014).sav"))
```

2. Split datasets:

```
# 1. Separate oh1_4 into rs1, rs2

oh2_2 <- oh1_4 %>%
  filter(rs_cohort == 2)

oh1_4 <- oh1_4 %>%
  filter(rs_cohort == 1)

# 2. Separate oh1_5 into rs1, rs2, rs3

oh3_2 <- oh1_5 %>%
  filter(rs_cohort == 3)

oh2_3 <- oh1_5 %>%
  filter(rs_cohort == 2)

oh1_5 <- oh1_5 %>%
  filter(rs_cohort == 1)

# 3. Separate oh1_6 into rs1 and rs2

oh1_6 <- oh1_6 %>%
  filter(rs_cohort == 1)
```

```
oh2_4 <- oh1_6 %>%
  filter(rs_cohort == 2)
```

3. Merge the data for all visits, by cohort:

```
### Merge RSI

oh1 <- list(oh1_1, oh1_2, oh1_3, oh1_4, oh1_5, oh1_6)

rs1_oh <- reduce(oh1, left_join, by = c("ergoid"))

### Merge RS2

oh2 <- list(oh2_1, oh2_2, oh2_3, oh2_4)

rs2_oh <- reduce(oh2, left_join, by = c("ergoid"))

### Merge RS3

oh3 <- list(oh3_1, oh3_2)

rs3_oh <- reduce(oh3, left_join, by = c("ergoid"))
```

4. Select and rename the specific variables from the combined dataset, by cohort:

```
rs1_oh <- rs1_oh %>%
  select(ergoid, antalc, e2_Alc_Tot, e3_Alc_Tot, e4_Alc_Tot, e5_Alc_tot, e6_alc_item_sum) %>%
  rename(oh1 = antalc,
         oh2 = e2_Alc_Tot,
         oh3 = e3_Alc_Tot,
         oh4 = e4_Alc_Tot,
         oh5 = e5_Alc_tot,
         oh6 = e6_alc_item_sum)

rs2_oh <- rs2_oh %>%
  select(ergoid, ep_Alc_Tot, e4_Alc_Tot, e5_Alc_tot,e6_alc_item_sum) %>%
  rename(
    oh3 = ep_Alc_Tot,
    oh4 = e4_Alc_Tot,
    oh5 = e5_Alc_tot,
    oh6 = e6_alc_item_sum)

rs3_oh <- rs3_oh %>%
  select(ergoid, ej_Alc_tot, e5_Alc_tot) %>%
  rename(
    oh4 = ej_Alc_tot,
    oh5 = e5_Alc_tot)
```

6. Bind , the cohorts. Since the variable names are consistent through the datasets.

```
alcohol <- rs1_oh %>%
  bind_rows(rs2_oh) %>%
  bind_rows(rs3_oh)
```

7. Export the variable to a `clean_data` folder.

```
export(alcohol, here::here("02_clean_data", "alcohol.Rdata"))
```

Chapter 9

Smoke anything

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Smoking

1. Import data for all cohorts

```
smoke1_1 <- read_sav(here::here("00_raw_data", "smoke", "e1_intvw_SMOKING_(23-nov-2011).sav"))
smoke1_2 <- read_sav(here::here("00_raw_data", "smoke", "e2_intvw_SMOKING_(23-nov-2011).sav"))
smoke1_3 <- read_sav(here::here("00_raw_data", "smoke", "e3_intvw_SMOKING_(11-nov-2011).sav"))
smoke1_4 <- read_sav(here::here("00_raw_data", "smoke", "e4_intvw_SMOKING_(04-nov-2011).sav"))
smoke1_5 <- read_sav(here::here("00_raw_data", "smoke", "e5_intvw_SMOKING_(04-sep-2014).sav"))
smoke2_1 <- read_sav(here::here("00_raw_data", "smoke", "ep_intvw_SMOKING_(30-sep-2011).sav"))
smoke3_1 <- read_sav(here::here("00_raw_data", "smoke", "ej_intvw_SMOKING_(28-mar-2011).sav"))
smoke1_6 <- read_sav(here::here("00_raw_data", "smoke", "e6_intvw_SMOKING_(20-feb-2017).sav"))
```

2. Split datasets:

```
# 1. Separate smoke1_4 into rs1, rs2

smoke2_2 <- smoke1_4 %>%
  filter(rs_cohort == 2)

smoke1_4 <- smoke1_4 %>%
  filter(rs_cohort == 1)

# 2. Separate smoke1_5 into rs1, rs2, rs3

smoke3_2 <- smoke1_5 %>%
  filter(rs_cohort == 3)

smoke2_3 <- smoke1_5 %>%
  filter(rs_cohort == 2)

smoke1_5 <- smoke1_5 %>%
  filter(rs_cohort == 1)

# 3. Separate smoke1_6 into rs1 and rs2

smoke1_6 <- smoke1_6 %>%
  filter(rs_cohort == 1)
```

```
smoke2_4 <- smoke1_6 %>%
  filter(rs_cohort == 2)
```

3. Merge the data for all visits, by cohort:

```
### Merge RSI

smoke1 <-
  list(smoke1_1, smoke1_2, smoke1_3, smoke1_4, smoke1_5, smoke1_6)

rs1_smoke <- reduce(smoke1, left_join, by = c("ergoid", "rs_cohort"))

### Merge RS2

smoke2 <- list(smoke2_1, smoke2_2, smoke2_3, smoke2_4)

rs2_smoke <-
  reduce(smoke2, left_join, by = c("ergoid", "rs_cohort"))

### Merge RS3

rs3_smoke <- smoke3_1 %>%
  left_join(smoke3_2, by = c("ergoid", "rs_cohort"))
```

4. Create variable smoke (0: never, 1: former, 2: current, NA) using the multiple questions related to smoking habits.

```
rs1_smoke <- rs1_smoke %>%
  mutate(
    smoke1 = ifelse((e1_ai7_20 == 0 & e1_ai7_30 == 0), 0, NA),
    smoke1 = ifelse((e1_ai7_20 == 0 &
                       e1_ai7_30 == 1), 1, smoke1),
    smoke1 = ifelse(e1_ai7_20 == 1, 2, smoke1),
    smoke2 = ifelse(e2_b0cg == 3 &
                      e2_b0pi == 3 & e2_b0ct == 3, 0, NA),
    smoke2 = ifelse((e2_b0cg == 2 |
                       e2_b0pi == 2 | e2_b0ct == 2), 1, smoke2),
    smoke2 = ifelse((e2_b0cg == 1 |
                       e2_b0pi == 1 | e2_b0ct == 1), 2, smoke2),
    smoke3 = ifelse(e3_cicg == 0 &
                      e3_cipi == 0 & e3_cict == 0 & e3_cictps == 0, 0, NA),
    smoke3 = ifelse((
      e3_cicg == 1 |
        e3_cipi == 1 | (e3_cict == 0 & e3_cictps == 1)
    ), 1, smoke3),
    smoke3 = ifelse((e3_cicg == 2 |
                       e3_cicg == 3 |
                       e3_cipi == 2 | e3_cipi == 3 | e3_cict == 1),
                    2,
                    smoke3
    ),
    smoke4 = ifelse(e4_dicg == 0 &
                      e4_dipi == 0 & e4_dict == 0, 0, NA),
    smoke4 = ifelse((
      e4_dicg == 1 |
```

```
        e4_dipi == 1 | e4_dict == 2 | e4_dict == 3
    ), 1, smoke4),
    smoke4 = ifelse((e4_dicg == 2 |
                       e4_dicg == 3 |
                       e4_dipi == 2 | e4_dipi == 3 | e4_dict == 1),
                    2,
                    smoke4
    ),
    smoke5 = ifelse(e5_EILF6 == 0 &
                      e5_EILFE == 0 & e5_EILF5 == 0 & e5_EILF4 == 0, 0, NA),
    smoke5 = ifelse(((e5_EILF6 == 0 &
                        e5_EILFE == 1) |
                       e5_EILF5 == 1 |
                       e5_EILF4 == 1 | e5_EILF4 == 2 | e5_EILF4 == 3
    ),
    1,
    smoke5
    ),
    smoke5 = ifelse((
      e5_EILF6 == 1 |
        e5_EILF5 == 2 |
        e5_EILF4 == 4 | e5_EILF4 == 5 | e5_EILF4 == 6
    ),
    2,
    smoke5
    ),
    smoke6 = ifelse((
      e6_EILF6 == 0 &
        e6_EILSE5 == 0 & e6_EILF5 == 0 & e6_EILF4 == 0
    ), 0, NA),
    smoke6 = ifelse(((e6_EILF6 == 0 &
                        (e6_EILSE5 == 1 |
                           e6_EILSE5 == 2)) |
                       e6_EILF5 == 1 |
                       e6_EILF4 == 1 | e6_EILF4 == 2 | e6_EILF4 == 3
    ),
    1,
    smoke6
    ),
    smoke6 = ifelse((
      e6_EILF6 == 1 |
        e6_EILF5 == 2 |
        e6_EILF4 == 4 | e6_EILF4 == 5 | e6_EILF4 == 6
    ),
    2,
    smoke6
    )
  ) %>%
  select(ergoid, smoke1, smoke2, smoke3, smoke4, smoke5)

rs2_smoke <- rs2_smoke %>%
  mutate(
    smoke3 = ifelse(ep_lf4 == 0 &
```

```
                    ep_lf5 == 0 & ep_lf6 == 0 & ep_lf6e == 0, 0, NA),
  smoke3 = ifelse((
    ep_lf4 == 1 |
      ep_lf5 == 1 | (ep_lf6 == 0 & ep_lf6e == 1)
  ), 1, smoke3),
  smoke3 = ifelse((ep_lf4 == 2 |
                     ep_lf5 == 2 | ep_lf6 == 1), 2, smoke3),
  smoke4 = ifelse(e4_dicg == 0 &
                    e4_dipi == 0 & e4_dict == 0, 0, NA),
  smoke4 = ifelse((
    e4_dicg == 1 |
      e4_dipi == 1 | e4_dict == 2 | e4_dict == 3
  ), 1, smoke4),
  smoke4 = ifelse((e4_dicg == 2 |
                     e4_dicg == 3 |
                     e4_dipi == 2 | e4_dipi == 3 | e4_dict == 1),
                  2,
                  smoke4
  ),
  smoke5 = ifelse(e5_EILF6 == 0 &
                    e5_EILFE == 0 & e5_EILF5 == 0 & e5_EILF4 == 0, 0, NA),
  smoke5 = ifelse(((e5_EILF6 == 0 &
                      e5_EILFE == 1) |
                     e5_EILF5 == 1 |
                     e5_EILF4 == 1 | e5_EILF4 == 2 | e5_EILF4 == 3
  ),
  1,
  smoke5
  ),
  smoke5 = ifelse((
    e5_EILF6 == 1 |
      e5_EILF5 == 1 |
      e5_EILF4 == 4 | e5_EILF4 == 5 | e5_EILF4 == 6
  ),
  2,
  smoke5
  ),
  smoke6 = ifelse(e6_EILF6 == 0 &
                    e6_EILSE5 == 0 & e6_EILF5 == 0 & e6_EILF4 == 0, 0, NA),
  smoke6 = ifelse(((e6_EILF6 == 0 &
                      (e6_EILSE5 == 1 |
                         e6_EILSE5 == 2)) |
                     e6_EILF5 == 1 |
                     e6_EILF4 == 1 | e6_EILF4 == 2 | e6_EILF4 == 3
  ),
  1,
  smoke6
  ),
  smoke6 = ifelse((
    e6_EILF6 == 1 |
      e6_EILF5 == 2 |
      e6_EILF4 == 4 | e6_EILF4 == 5 | e6_EILF4 == 6
  ),
```

```
    2,
    smoke6
    )
  ) %>%
  select(ergoid, rs_cohort, smoke3, smoke4, smoke5, smoke6)


rs3_smoke <- rs3_smoke %>%
  mutate(
    smoke4 = ifelse(ej_yilf4 == 0 &
                      ej_yilf5 == 0 & ej_yilf6 == 0 & ej_yilfe == 0, 0, NA),
    smoke4 = ifelse((
      ej_yilf4 == 1 |
        ej_yilf4 == 2 |
        ej_yilf4 == 3 |
        ej_yilf5 == 1 | (ej_yilf6 == 0 & ej_yilfe == 1)
    ),
    1,
    smoke4
    ),
    smoke4 = ifelse((
      ej_yilf4 == 4 |
        ej_yilf4 == 5 |
        ej_yilf4 == 6 | ej_yilf5 == 2 | ej_yilf6 == 1
    ),
    2,
    smoke4
    ),
    smoke5 = ifelse(e5_EILF6 == 0 &
                      e5_EILFE == 0 & e5_EILF5 == 0 & e5_EILF4 == 0, 0, NA),
    smoke5 = ifelse(((e5_EILF6 == 0 &
                        e5_EILFE == 1) |
                       e5_EILF5 == 1 |
                       e5_EILF4 == 1 | e5_EILF4 == 2 | e5_EILF4 == 3
    ),
    1,
    smoke5
    ),
    smoke5 = ifelse((
      e5_EILF6 == 1 |
        e5_EILF5 == 1 |
        e5_EILF4 == 4 | e5_EILF4 == 5 | e5_EILF4 == 6
    ),
    2,
    smoke5
    )
  ) %>%
  select(ergoid, rs_cohort, smoke4, smoke5)
```

6. Bind , the cohorts. Since the variable names are consistent through the datasets.

```
smoke <- rs1_smoke %>%
  bind_rows(rs2_smoke) %>%
  bind_rows(rs3_smoke)
```

7. Export the variable to a `clean_data` folder.

```
export(smoke, here::here("02_clean_data", "smoke.Rdata"))
```

Chapter 10

Smoke cigarettes

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Smoking

1. Import data for all cohorts

```
smoke1_1 <- read_sav(here::here("00_raw_data", "smoke", "e1_intvw_SMOKING_(23-nov-2011).sav"))
smoke1_2 <- read_sav(here::here("00_raw_data", "smoke", "e2_intvw_SMOKING_(23-nov-2011).sav"))
smoke1_3 <- read_sav(here::here("00_raw_data", "smoke", "e3_intvw_SMOKING_(11-nov-2011).sav"))
smoke1_4 <- read_sav(here::here("00_raw_data", "smoke", "e4_intvw_SMOKING_(04-nov-2011).sav"))
smoke1_5 <- read_sav(here::here("00_raw_data", "smoke", "e5_intvw_SMOKING_(04-sep-2014).sav"))
smoke2_1 <- read_sav(here::here("00_raw_data", "smoke", "ep_intvw_SMOKING_(30-sep-2011).sav"))
smoke3_1 <- read_sav(here::here("00_raw_data", "smoke", "ej_intvw_SMOKING_(28-mar-2011).sav"))
smoke1_6 <- read_sav(here::here("00_raw_data", "smoke", "e6_intvw_SMOKING_(20-feb-2017).sav"))
```

2. Split datasets (same steps as in @??smoke):

```
# 1. Separate smoke1_4 into rs1, rs2

smoke2_2 <- smoke1_4 %>%
  filter(rs_cohort == 2)

smoke1_4 <- smoke1_4 %>%
  filter(rs_cohort == 1)

# 2. Separate smoke1_5 into rs1, rs2, rs3

smoke3_2 <- smoke1_5 %>%
  filter(rs_cohort == 3)

smoke2_3 <- smoke1_5 %>%
  filter(rs_cohort == 2)

smoke1_5 <- smoke1_5 %>%
  filter(rs_cohort == 1)

# 3. Separate smoke1_6 into rs1 and rs2

smoke1_6 <- smoke1_6 %>%
  filter(rs_cohort == 1)
```

```
smoke2_4 <- smoke1_6 %>%
  filter(rs_cohort == 2)
```

3. Merge the data for all visits, by cohort:

```
### Merge RSI

smoke1 <-
  list(smoke1_1, smoke1_2, smoke1_3, smoke1_4, smoke1_5, smoke1_6)

rs1_smoke<- reduce(smoke1, left_join, by = c("ergoid", "rs_cohort"))

### Merge RS2

smoke2 <- list(smoke2_1, smoke2_2, smoke2_3, smoke2_4)

rs2_smoke <-
  reduce(smoke2, left_join, by = c("ergoid", "rs_cohort"))

### Merge RS3

rs3_smoke <- smoke3_1 %>%
  left_join(smoke3_2, by= c("ergoid", "rs_cohort"))
```

4. Create variable `smoke_cig` (0: never, 1: former, 2: current, NA) using only the questions related to smoking cigarettes.

```
rs1_cig <- rs1_smoke %>%
  mutate(
    smoke_cig1 = ifelse((e1_ai7_20 == 0 & e1_ai7_30 == 0), 0, NA),
    smoke_cig1 = ifelse((e1_ai7_20 == 0 &
                           e1_ai7_30 == 1), 1, smoke_cig1),
    smoke_cig1 = ifelse(e1_ai7_20 == 1, 2, smoke_cig1),
    smoke_cig2 = ifelse(e2_b0ct == 3, 0, NA),
    smoke_cig2 = ifelse(e2_b0ct == 2, 1, smoke_cig2),
    smoke_cig2 = ifelse(e2_b0ct == 1, 2, smoke_cig2),
    smoke_cig3 = ifelse((e3_cict == 0 &
                           e3_cictps == 0), 0, NA),
    smoke_cig3 = ifelse((e3_cict == 0 &
                           e3_cictps == 1), 1, smoke_cig3),
    smoke_cig3 = ifelse(e3_cict == 1, 2, smoke_cig3),
    smoke_cig4 = ifelse(e4_dict == 0, 0, NA),
    smoke_cig4 = ifelse((e4_dict == 2 |
                           e4_dict == 3), 1, smoke_cig4),
    smoke_cig4 = ifelse(e4_dict == 1, 2, smoke_cig4),
    smoke_cig5 = ifelse(e5_EILF6 == 0 & e5_EILFE == 0, 0, NA),
    smoke_cig5 = ifelse((e5_EILF6 == 0 &
                           e5_EILFE == 1), 1, smoke_cig5),
    smoke_cig5 = ifelse(e5_EILF6 == 1, 2, smoke_cig5),
    smoke_cig6 = ifelse((e6_EILF6 == 0 &
                           e6_EILSE5 == 0), 0, NA),
    smoke_cig6 = ifelse((
      e6_EILF6 == 0 & (e6_EILSE5 == 1 | e6_EILSE5 == 2)
    ), 1, smoke_cig6),
    smoke_cig6 = ifelse((e6_EILF6 == 1), 2, smoke_cig6)
```

```
  ) %>%
  select(ergoid, starts_with("smoke_cig"))

rs2_cig <- rs2_smoke %>%
  mutate(
    smoke_cig3 = ifelse((ep_lf6 == 0 & ep_lf6e == 0), 0, NA),
    smoke_cig3 = ifelse((ep_lf6 == 0 &
                           ep_lf6e == 1), 1, smoke_cig3),
    smoke_cig3 = ifelse(ep_lf6 == 1, 2, smoke_cig3),
    smoke_cig4 = ifelse(e4_dict == 0, 0, NA),
    smoke_cig4 = ifelse((e4_dict == 2 |
                           e4_dict == 3), 1, smoke_cig4),
    smoke_cig4 = ifelse(e4_dict == 1, 2, smoke_cig4),
    smoke_cig5 = ifelse(e5_EILF6 == 0 & e5_EILFE == 0, 0, NA),
    smoke_cig5 = ifelse((e5_EILF6 == 0 &
                           e5_EILFE == 1), 1, smoke_cig5),
    smoke_cig5 = ifelse(e5_EILF6 == 1, 2, smoke_cig5),
    smoke_cig6 = ifelse((e6_EILF6 == 0 &
                           e6_EILSE5 == 0), 0, NA),
    smoke_cig6 = ifelse((
      e6_EILF6 == 0 & (e6_EILSE5 == 1 | e6_EILSE5 == 2)
    ), 1, smoke_cig6),
    smoke_cig6 = ifelse((e6_EILF6 == 1), 2, smoke_cig6)
  ) %>%
  select(ergoid, starts_with("smoke_cig"))

rs3_cig <- rs3_smoke %>%
  mutate(
    smoke_cig4 = ifelse(ej_yilf6 == 0 & ej_yilfe == 0, 0, NA),
    smoke_cig4 = ifelse(ej_yilf6 == 0 &
                          ej_yilfe == 1, 1, smoke_cig4),
    smoke_cig4 = ifelse(ej_yilf6 == 1, 2, smoke_cig4),
    smoke_cig5 = ifelse((e5_EILF6 == 0 &
                           e5_EILFE == 0), 0, NA),
    smoke_cig5 = ifelse((e5_EILF6 == 0 &
                           e5_EILFE == 1), 1, smoke_cig5),
    smoke_cig5 = ifelse(e5_EILF6 == 1, 2, smoke_cig5)
  ) %>%
  select(ergoid, smoke_cig4, smoke_cig5)
```

6. Bind , the cohorts. Since the variable names are consistent through the datasets.

```
smoke_cig <- rs1_cig %>%
  bind_rows(rs2_cig) %>%
  bind_rows(rs3_cig)
```

7. Export the variable to a `clean_data` folder.

```
export(smoke_cig, here::here("02_clean_data", "smoke_cig.Rdata"))
```

Chapter 11

Hypertension

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Hypertension

1. Import data for all cohorts

```
ht <- read_sav(here::here("00_raw_data", "hypertension", "HT2018_analysisfile_(15-may-2018).sav"))
```

2. Split datasets:

```
# 1. Separate hypertension for RS-I, RS-II and RS-III

ht1 <- ht %>%
  filter(rs_cohort == 1)

ht2 <- ht %>%
  filter(rs_cohort == 2)

ht3 <- ht %>%
  filter(rs_cohort == 3)
```

3. Select and rename variables:

```
ht1 <- ht %>%
  select(ergoid, contains("systolicBP"), contains("diastolicBP"), contains("HT2018"), contains("bpldrug
  rename(
    sbp1 = e1_systolicBP,
    sbp2 = e2_systolicBP,
    sbp3 = e3_systolicBP,
    sbp4 = e4_systolicBP,
    sbp5 = e5_systolicBP,
    sbp6 = e6_systolicBP,
    dbp1 = e1_diastolicBP,
    dbp2 = e2_diastolicBP,
    dbp3 = e3_diastolicBP,
    dbp4 = e4_diastolicBP,
    dbp5 = e5_diastolicBP,
    dbp6 = e6_diastolicBP,
    ht1 = e1_HT2018,
    ht2 = e2_HT2018,
    ht3 = e3_HT2018,
    ht4 = e4_HT2018,
```

```
    ht5 = e5_HT2018,
    ht6 = e6_HT2018,
    htdrug1 = e1_bpldrug,
    htdrug2 = e2_bpldrug,
    htdrug3 = e3_bpldrug,
    htdrug4 = e4_bpldrug,
    htdrug5 = e5_bpldrug,
    htdrug6 = e6_bpldrug
    )


ht2 <- ht %>%
  select(ergoid, contains("systolicBP"), contains("diastolicBP"), contains("HT2018"), contains("bpldrug
  rename(
    sbp3 = ep_systolicBP,
    sbp4 = e4_systolicBP,
    sbp5 = e5_systolicBP,
    sbp6 = e6_systolicBP,
    dbp3 = ep_diastolicBP,
    dbp4 = e4_diastolicBP,
    dbp5 = e5_diastolicBP,
    dbp6 = e6_diastolicBP,
    ht3 = ep_HT2018,
    ht4 = e4_HT2018,
    ht5 = e5_HT2018,
    ht6 = e6_HT2018,
    htdrug3 = ep_bpldrug,
    htdrug4 = e4_bpldrug,
    htdrug5 = e5_bpldrug,
    htdrug6 = e6_bpldrug
    )

ht3 <- ht %>%
  filter(rs_cohort == 3) %>%
  select(ergoid, contains("systolicBP"), contains("HT2018")) %>%
  rename(
    sbp4 = ej_systolicBP,
    sbp5 = e5_systolicBP,
    ht4 = ej_HT2018,
    ht5 = e5_HT2018)
```

6. Bind , the cohorts. Since the variable names are consistent through the datasets.

```
smoke_cig <- rs1_cig %>%
  bind_rows(rs2_cig) %>%
  bind_rows(rs3_cig)
```

7. Export the variable to a `clean_data` folder.

```
export(smoke_cig, here::here("02_clean_data", "smoke_cig.Rdata"))
```

Chapter 12

Laboratory data

Source: https://epi-wiki.erasmusmc.nl/wiki/ergowiki/index.php/Cholesterol

We included cholesterol and hdl since they were sistematically collected across visits for all cohorts. *Note*: There are no laboratory data for visit RS-I-2

Glucose has the subfix: **v3846**, but since it was collected in different ways and not in all waves, we don't include it, but it could be collected using the same code.

1. Import data for all cohorts

```
chol1_1 <- import(here::here("00_raw_data", "cholesterol",  "e1_CHOLESTEROL_(10.03.2010).sav"))
chol1_3 <- import(here::here("00_raw_data", "cholesterol",  "e3_(3)_LAB_(10-mar-2010).sav"))
chol1_4 <- import(here::here("00_raw_data", "cholesterol",  "e4_(4)_LAB_(10-mar-2010)b.sav"))
chol1_5 <- import(here::here("00_raw_data", "cholesterol",  "e5_(5)_LAB_(29-aug-2014)r.sav"))
chol2_1 <- import(here::here("00_raw_data", "cholesterol",  "ep_(1)_LAB_(15-mar-2010).sav"))
chol3_1 <- import(here::here("00_raw_data", "cholesterol",  "ej_(1)_LAB_(11-jun-2009)r.sav"))
```

2. Split datasets:

```
# 1. Separate chol1_4 into rs1, rs2, rs3

chol2_2 <- chol1_4 %>%
  filter(rs_cohort == 2)

chol1_4 <- chol1_4 %>%
  filter(rs_cohort == 1)

# 2. Separate chol1_5 into rs1, rs2, rs3

chol3_2 <- chol1_5 %>%
  filter(rs_cohort == 3)

chol2_3 <- chol1_5 %>%
  filter(rs_cohort == 2)

chol1_5 <- chol1_5 %>%
  filter(rs_cohort == 1)

# 3. Separate chol1_6 into rs1 and rs2
```

```
# Not available
```

3. Merge the data for all visits, by cohort:

```
### Merge RSI

chol1 <- list(chol1_1, chol1_3, chol1_4, chol1_5)#, chol1_6)

rs1_chol <- reduce(chol1, left_join, by = c("ergoid"))

### Merge RSII

chol2 <- list(chol2_1, chol2_2, chol2_3)#,chol2_4)

rs2_chol <- reduce(chol2, left_join, by = c("ergoid"))

### Merge RSIII

chol3 <- list(chol3_1, chol3_2)

rs3_chol <- reduce(chol3, left_join, by = c("ergoid"))
```

4. Select and rename variables:

```
rs1_chol <- rs1_chol %>%
  select(ergoid, e1_al7_chl, e1_al7_hdl, ends_with("_3845"), ends_with("_4107")) %>%
  rename(chol1 = e1_al7_chl,
         chol3 = e3_3845,
         chol4 = e3_3845,
         chol5 = e5_3845,
         hdl1 = e1_al7_hdl,
         hdl3 = e3_4107,
         hdl4 = e4_4107,
         hdl5 = e5_4107)

rs2_chol <- rs2_chol %>%
  select(ergoid, ends_with("_3845"), ends_with("_4107")) %>%
  rename(chol3 = ep_3845,
         chol4 = e4_3845,
         chol5 = e5_3845,
         hdl3 = ep_4107,
         hdl4 = e4_4107,
         hdl5 = e5_4107)


rs3_chol <- rs3_chol %>%
  select(ergoid, ends_with("_3845"), ends_with("_4107")) %>%
  rename(chol4 = ej_3845,
         chol5 = e5_3845,
         hdl4 = ej_4107,
         hdl5 = e5_4107)
```

5. Bind , the cohorts. Since the variable names are consistent through the datasets.

```
chol <- rs1_chol %>%
  bind_rows(rs2_chol) %>%
```

```
  bind_rows(rs3_chol)
```

6. Export the variable to a `clean_data` folder.

```
export(chol, here::here("02_clean_data", "chol.Rdata"))
```