

# Double Check Clinical vars

This calculations were run (on a kimmel lab machine) to double check the coding of cases and controls and Sex in our phenotypic file.

While most things are fine. Two issues were identified:

1. Two participants in COBRE are coded as Controls, but appear to be cases
2. 30 (SSD) participants from ZHH look like they might be repeated.

```
library(dplyr)
library(tidyr)
library(readr)
library(stringr)
library(ggplot2)
library(knitr)
```

## Reading Saba's clinical data page

```
## reading in the qced_sublists csv to get the sublists
qced_sublists <- read_csv("../phenotypic/NEWallSubjects_completeData3_DM_not_sexmatched.csv",
                           col_types = c(
                             X1 = col_integer(),
                             name = col_character(),
                             subid = col_character(),
                             DX_GROUP = col_integer(),
                             mean_fd = col_double(),
                             age = col_integer(),
                             sex.x = col_integer(),
                             educationCode = col_double(),
                             site = col_integer(),
                             sex.y = col_integer(),
                             X.bad_fd = col_double(),
                             global_corr = col_double(),
                             mean_snfr = col_double()
                           ))

## Warning: Missing column names filled in: 'X1' [1]

pheno <- qced_sublists %>%
  mutate(Site = factor(site, levels = c(1,2,3),
                                labels = c("CMH","ZHH","COBRE")) ,
         DX = factor(DX_GROUP, level = c(1,2), labels = c('SSD', 'Ctrl')),
         Sex = factor(sex.x, levels = c(1,2), labels = c('M','F')),
         Edu = if_else(is.na(educationCode),
                       true = mean(educationCode, na.rm = T),
                       false = educationCode))
```

## reading the COBRE demographics page

```
schizconnect_COBRE_assessmentData_3125 <- read_csv("/external/SchizConnect/COBRE/schizconnect_COBRE_assessmentData_3125.csv",
                                                  col_types = c(
```

```

source = col_character(),
study = col_character(),
site = col_character(),
subjectid = col_character(),
visit = col_character(),
assessment = col_character(),
assessment_description = col_character(),
question_id = col_character(),
question_value = col_character()
)) %>%
mutate(subid = str_c('COBRE_', subjectid, '_SESS01'))

```

```

cobre_check <- inner_join(schizconnect_COBRE_assessmentData_3125, pheno, by = "subid")

```

```

cobre_check %>%
  filter(question_id == "CODEM_2") %>%
  select(sex.x, question_value) %>%
  ftable()

```

```

##      question_value Female Male
## sex.x
## 1                0    57
## 2                21     0

```

```

cobre_check %>%
  filter(question_id == "CODEM_17") %>%
  select(DX_GROUP, question_value) %>%
  ftable()

```

```

##      question_value 11 13 14 15 16 17 18 19 20 21 22 24 25 26 9999 md N/A
## DX_GROUP
## 1                1  0  1  2  1  3  5  2  0  3  2  1  1  1   0  0  0
## 2                0  1  0  0  0  0  0  0  1  0  0  0  0  0  14  1  14

```

```

cobre_weird <- cobre_check %>%
  filter(question_id == "CODEM_17",
         question_value %in% c("13", "20")) %>%
  select(subid)

inner_join(cobre_weird, cobre_check, by = "subid") %>%
  select(question_id, question_value, age, Sex, Edu, DX)

```

```

## # A tibble: 626 x 6
##   question_id question_value   age Sex   Edu DX
##   <chr>      <chr>         <int> <fct> <dbl> <fct>
## 1 CODEM_1    38             38 M    12.0 Ctrl
## 2 CODEM_10   Right            38 M    12.0 Ctrl
## 3 CODEM_11   no              38 M    12.0 Ctrl
## 4 CODEM_12   <NA>            38 M    12.0 Ctrl
## 5 CODEM_13   <NA>            38 M    12.0 Ctrl
## 6 CODEM_14   no              38 M    12.0 Ctrl
## 7 CODEM_15   <NA>            38 M    12.0 Ctrl
## 8 CODEM_16   20              38 M    12.0 Ctrl
## 9 CODEM_17   20              38 M    12.0 Ctrl
## 10 CODEM_18  20              38 M    12.0 Ctrl
## # ... with 616 more rows

```

Conclusion: These two participants are NOT Controls...

COBRE\_A00028189\_SESS01

COBRE\_A00035836\_SESS01

```
pheno_corrected <- pheno
for (badsub in cobre_weird$subid) {
  pheno_corrected$DX[pheno_corrected$subid == badsub] <- 'SSD'
  pheno_corrected$DX_GROUP[pheno_corrected$subid == badsub] <- 1
}

cobre_recheck <- inner_join(schizconnect_COBRE_assessmentData_3125, pheno_corrected, by = "subid")

cobre_recheck %>%
  filter(question_id == "CODEM_17") %>%
  select(DX_GROUP, question_value) %>%
  ftable()
```

```
##          question_value 11 13 14 15 16 17 18 19 20 21 22 24 25 26 9999 md N/A
## DX_GROUP
## 1                1  1  1  2  1  3  5  2  1  3  2  1  1  1    0  0  0
## 2                0  0  0  0  0  0  0  0  0  0  0  0  0  0    14  1  14
```

Now looking at ZHH data

```
library(readxl)
qrySZ_Sess_Miklos <- read_excel("/external/miklos/demographics/qrySZ_Sess_Miklos.xlsx")

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F2 / R2C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F3 / R3C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F4 / R4C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F5 / R5C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F6 / R6C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F7 / R7C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F8 / R8C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F9 / R9C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F10 / R10C6: got a date
## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F11 / R11C6: got a date
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

```

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F495 / R495C6: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F496 / R496C6: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F497 / R497C6: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F498 / R498C6: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F499 / R499C6: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F500 / R500C6: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F501 / R501C6: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in F502 / R502C6: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in H503 / R503C8: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in H504 / R504C8: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in H505 / R505C8: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in H506 / R506C8: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in H507 / R507C8: got a date

## Warning in read_fun(path = path, sheet = sheet, limits = limits, shim =
## shim, : Expecting numeric in H508 / R508C8: got a date

qryHC_Sess_Miklos <- read_excel("/external/miklos/demographics/qryHC_NewResting_15-40.xlsx")

pheno %>% filter(Site == "ZHH") %>% select(subid) %>% slice(1:5)

## # A tibble: 5 x 1
##   subid
##   <chr>
## 1 EXP_21359_SESS01
## 2 EXP_21362_SESS01
## 3 EXP_21366_SESS01
## 4 EXP_21367_SESS01
## 5 EXP_21368_SESS01

ZHH_HC <- qryHC_Sess_Miklos %>%
  mutate(subid = str_c('EXP_', SessNo, '_SESS01')) %>%
  inner_join(qced_sublists, by = "subid")

ZHH_HC %>%
  select(sex, sex.x) %>%
  ftable

```

```
##      sex.x  1  2
## sex
## 1          50  0
## 2           0 54
```

```
ZHH_HC %>%
  select(sex, DX_GROUP) %>%
  ftable
```

```
##      DX_GROUP  2
## sex
## 1          50
## 2          54
```

```
ZHH_SZ <- qrySZ_Sess_Miklos %>%
  mutate(subid = str_c('EXP_', SessNo, '_SESS01')) %>%
  inner_join(qced_sublists, by = "subid")
```

```
ZHH_SZ %>%
  select(sex, DX_GROUP) %>%
  ftable
```

```
##      DX_GROUP  1
## sex
## 1          100
## 2           31
```

```
ZHH_SZ %>%
  select(sex, sex.x) %>%
  ftable
```

```
##      sex.x  1  2
## sex
## 1          100  0
## 2           0 31
```

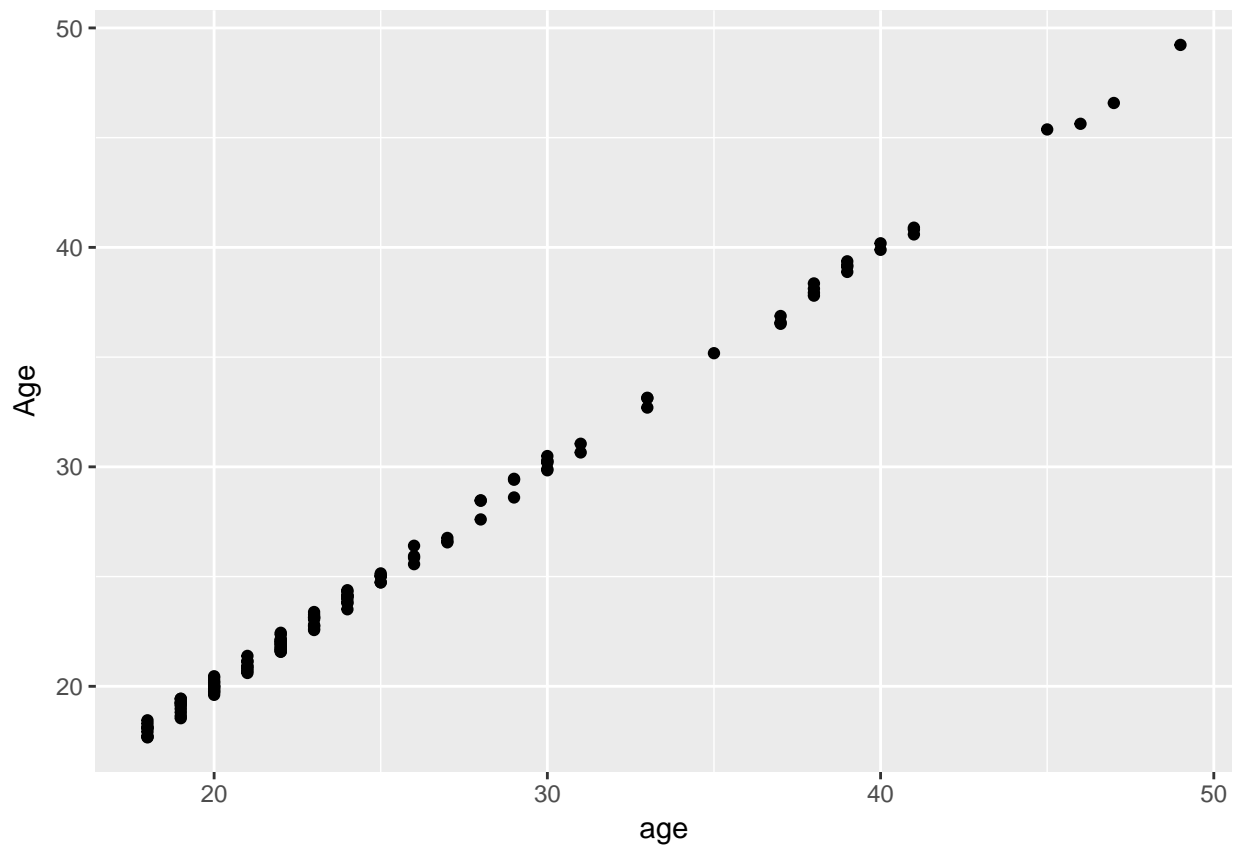
```
ZHH_SZ %>%
  select(sex, DX_GROUP) %>%
  ftable
```

```
##      DX_GROUP  1
## sex
## 1          100
## 2           31
```

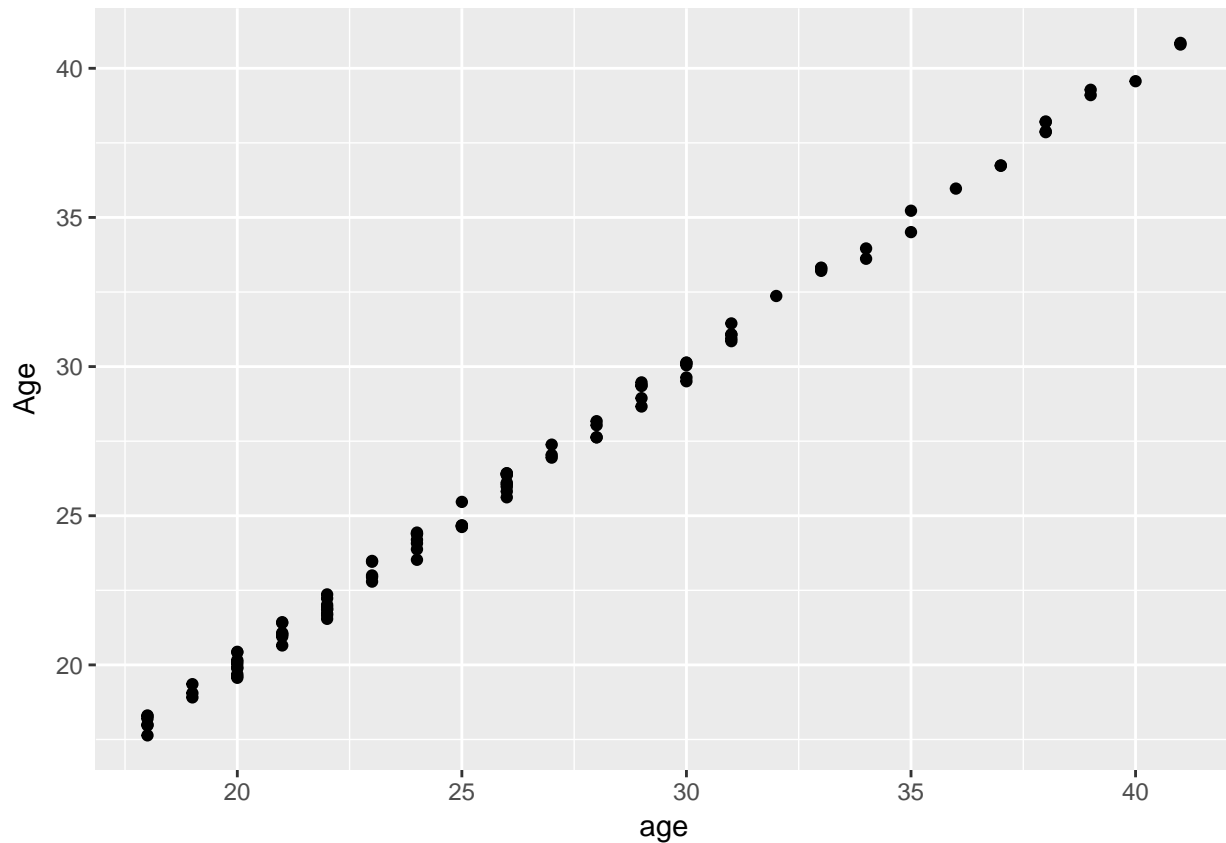
```
repeated_SZ_subjects <- ZHH_SZ %>%
  count(grid) %>%
  filter(n > 1)
```

```
repeated_HC_subjects <- ZHH_HC %>%
  count(GRID) %>%
  filter(n > 1)
```

```
ggplot(ZHH_SZ, aes(x = age, y = Age)) + geom_point()
```



```
ggplot(ZHH_HC, aes(x = age, y = Age)) + geom_point()
```

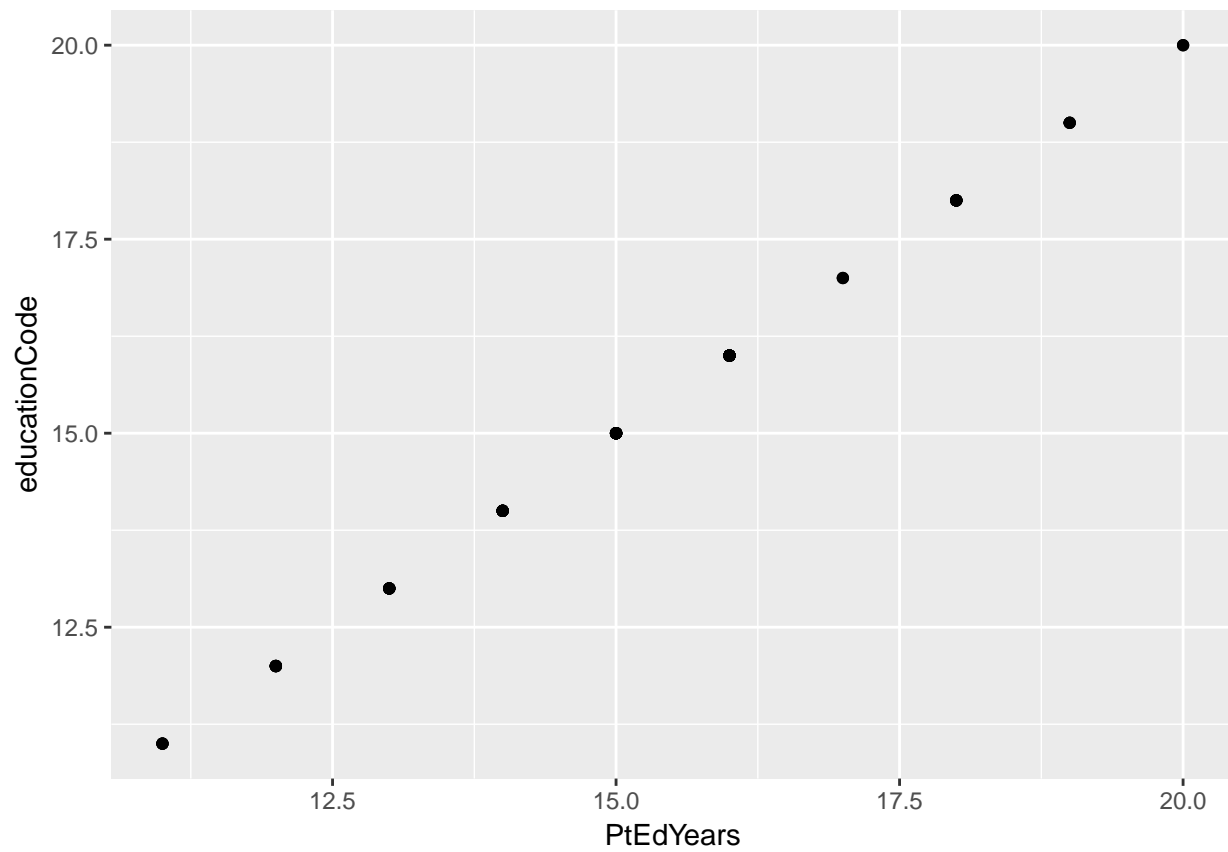


```
names(ZHH_HC)
```

```
## [1] "GRID" "sex" "PtEdYears"
## [4] "PtEdHighestHol_Demo" "SessNo" "STUDYID"
## [7] "ID" "Age" "SubjectType"
## [10] "NewRSfMRI" "EXAMTYPE" "EXAMDATE"
## [13] "subid" "X1" "name"
## [16] "DX_GROUP" "mean_fd" "age"
## [19] "sex.x" "educationCode" "site"
## [22] "sex.y" "X.bad_fd" "global_corr"
## [25] "mean_snfr"
```

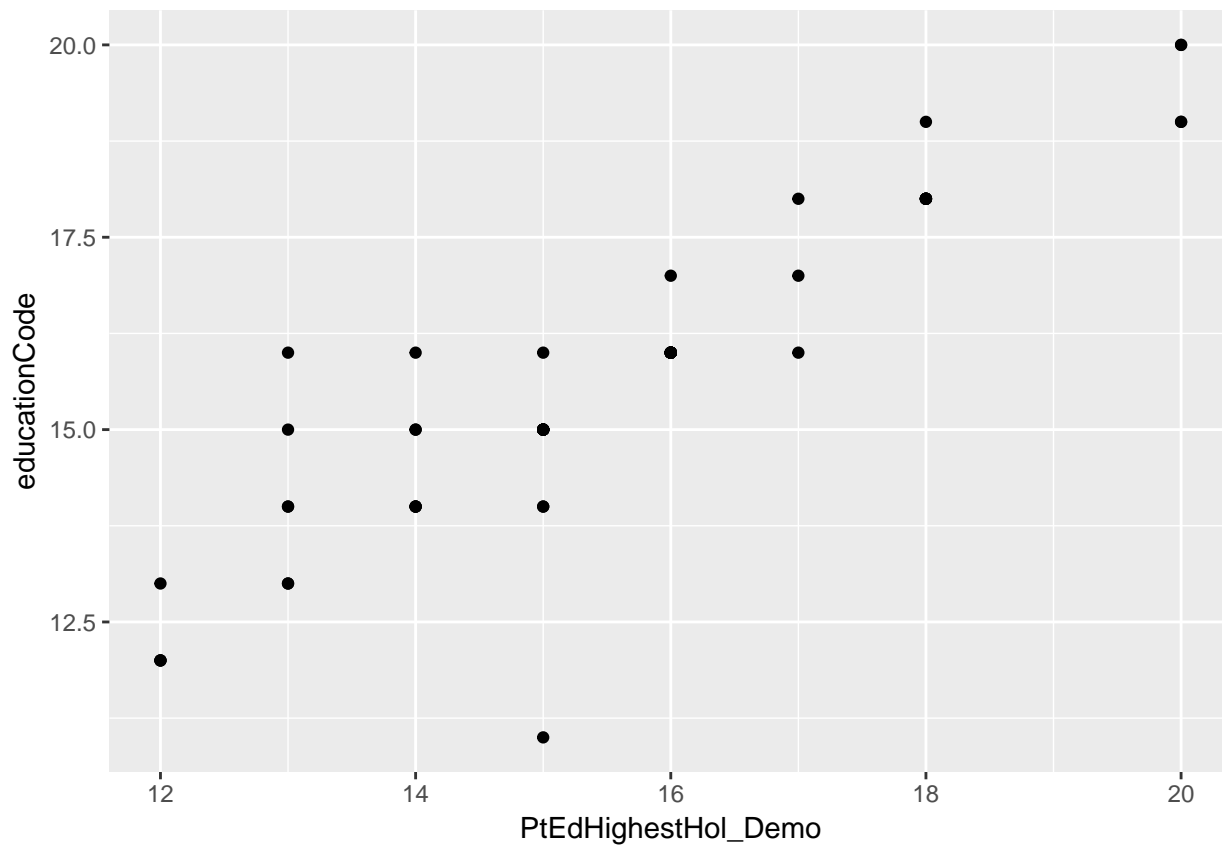
```
ggplot(ZHH_HC, aes(x = PtEdYears, y = educationCode)) + geom_point()
```

```
## Warning: Removed 31 rows containing missing values (geom_point).
```



```
ggplot(ZHH_HC, aes(x = PtEdHighestHol_Demo, y = educationCode)) + geom_point()
```

```
## Warning: Removed 42 rows containing missing values (geom_point).
```



```
names(ZHH_SZ)
```

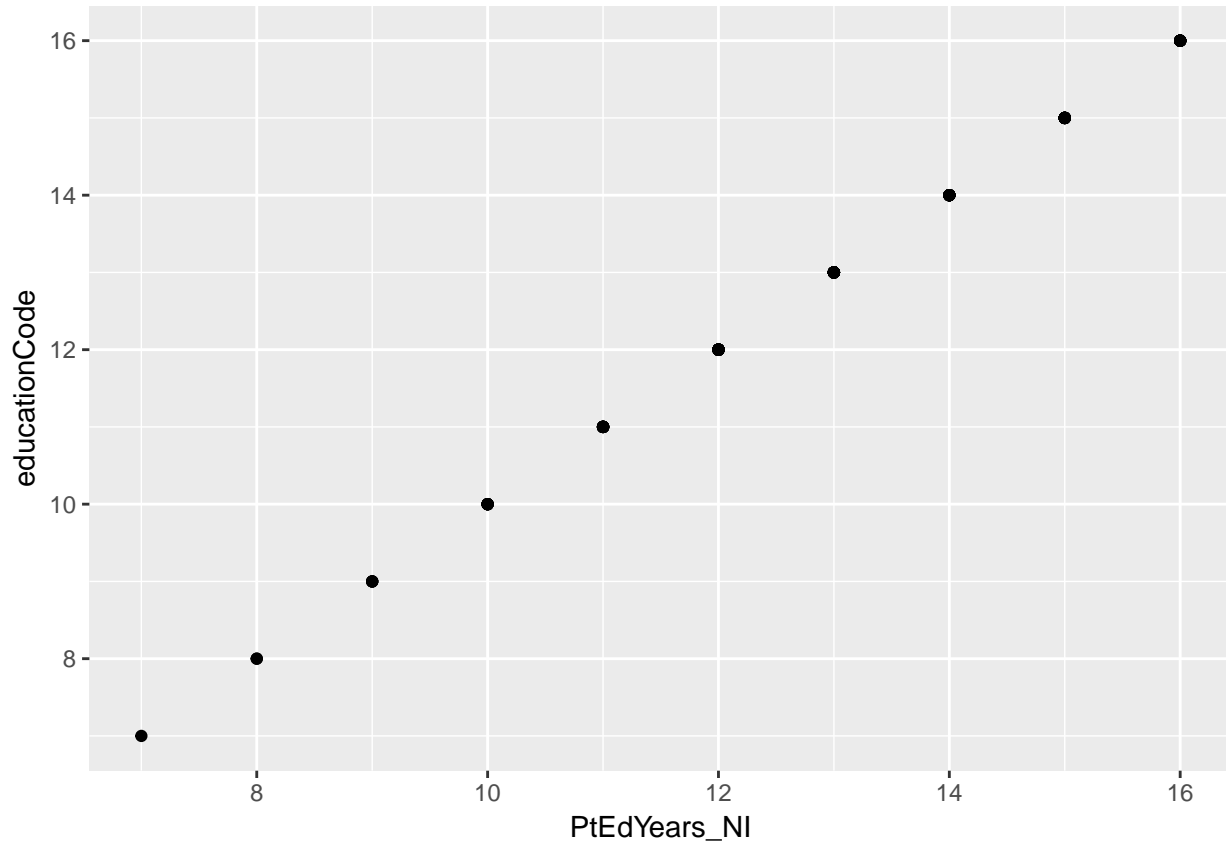
```
## [1] "grid"
## [2] "sex"
## [3] "race"
## [4] "PtEdYears_NI"
## [5] "PtEdHighestHol_Demo"
## [6] "Age"
## [7] "MRI_Study"
## [8] "DOS"
## [9] "SessNo"
## [10] "EXAMTYPE"
## [11] "Schizoaffective Disorder"
## [12] "Schizophrenia, Catatonic Type"
## [13] "Schizophrenia, Disorganized Type"
## [14] "Schizophrenia, Paranoid Type"
## [15] "Schizophrenia, Residual Type"
## [16] "Schizophrenia, Undifferentiated Type"
## [17] "Schizophreniform Disorder"
## [18] "subid"
## [19] "X1"
## [20] "name"
## [21] "DX_GROUP"
## [22] "mean_fd"
## [23] "age"
## [24] "sex.x"
## [25] "educationCode"
```



```
## [26] "site"  
## [27] "sex.y"  
## [28] "X.bad_fd"  
## [29] "global_corr"  
## [30] "mean_snfr"
```

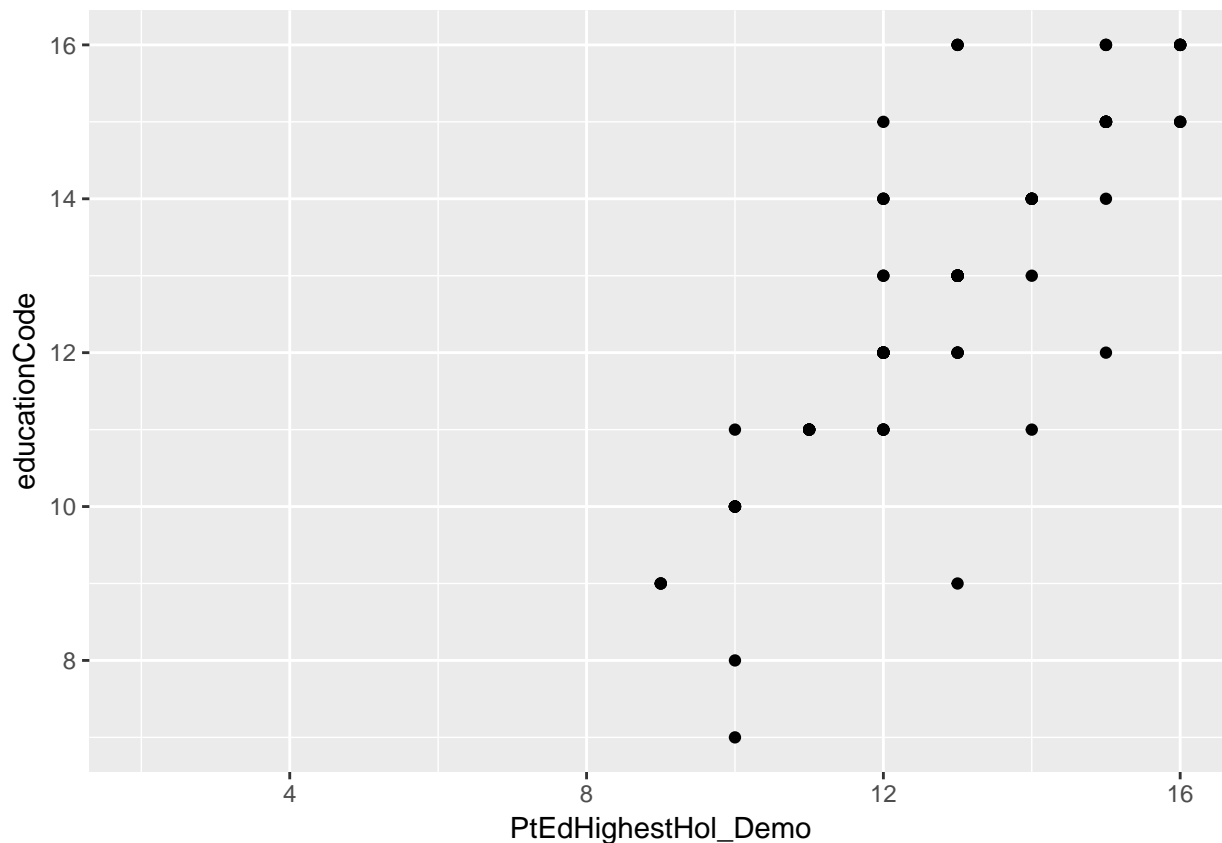
```
ggplot(ZHH_SZ, aes(x = PtEdYears_NI, y = educationCode)) + geom_point()
```

```
## Warning: Removed 14 rows containing missing values (geom_point).
```



```
ggplot(ZHH_SZ, aes(x = PtEdHighestHol_Demo, y = educationCode)) + geom_point()
```

```
## Warning: Removed 26 rows containing missing values (geom_point).
```



```
set.seed(14)
ZHH_SZ_subsampled <- qrySZ_Sess_Miklos %>%
  mutate(subid = str_c('EXP_', SessNo, '_SESS01')) %>%
  ungroup() %>%
  group_by(grid) %>%
  sample_n(1) %>%
  select(subid) %>%
  ungroup()
```

```
## Adding missing grouping variables: `grid`
```

```
to_drop <- ZHH_SZ %>%
  anti_join(ZHH_SZ_subsampled, by = "subid")

pheno_corrected1 <- pheno_corrected %>%
  anti_join(to_drop, by = "subid")
```

Now for SPINS

```
spins_demographics <- read_csv("/archive/data-2.0/SPINS/data/clinical/demographics.csv") %>%
  mutate(subid = str_c(record_id, '_01'))
```

```
## Parsed with column specification:
## cols(
##   record_id = col_character(),
##   group = col_character(),
```

```
## sex = col_integer(),
## age = col_integer(),
## education = col_integer(),
## terminated = col_integer()
## )

spins_check <- inner_join(spins_demographics, qced_sublists, by = "subid")

spins_demographics$subid[1:5]
```

```
## [1] "SPN01_CMH_0001_01" "SPN01_CMH_0002_01" "SPN01_CMH_0003_01"
## [4] "SPN01_CMH_0004_01" "SPN01_CMH_0005_01"
```

```
spins_check %>%
  select(group, DX_GROUP) %>%
  ftable()
```

```
##           DX_GROUP  1  2
## group
## case_arm_2         40  0
## control_arm_1        0 27
```

```
spins_check %>%
  select(sex, sex.x, group) %>%
  ftable()
```

```
##           group case_arm_2 control_arm_1
## sex sex.x
## 1  1           0           0
##    2          14          14
## 2  1          26          13
##    2           0           0
```

## Finally writting results from corrected version back to csv

```
names(pheno_corrected1)
```

```
## [1] "X1"           "name"         "subid"        "DX_GROUP"
## [5] "mean_fd"      "age"          "sex.x"        "educationCode"
## [9] "site"         "sex.y"        "X.bad_fd"     "global_corr"
## [13] "mean_snfr"    "Site"         "DX"           "Sex"
## [17] "Edu"
```

```
pheno_corrected1 %>%
  select(subid, name, Site, DX, age, Sex, Edu, mean_fd, X.bad_fd, global_corr, mean_snfr) %>%
  write_csv('./phenotypic/subjects_not_sexmatched_20180507.csv')
```