Supplement 1

Demographic characteristics

Peter Kamerman

Last knitted: 27 June 2019

Contents

Import data **1**

Quick look **1**

Check missingness **2**
Full cohort . . . 2
HIV- . . . 2
HIV+ . . . 3

Numeric data **3**
Age . . . 3
Point estimates . . . 3
Full cohort . . . 3
By HIV status . . . 4
95% CI of the point estimates . . . 6
Whole cohort . . . 6
By HIV status . . . 6
95% CI of the difference in mean . . . 6
CD4 T-cell count . . . 8
Point estimates . . . 8
95% CI of the point estimates . . . 9

Hopkins Symptom Checklist 25 **10**
Depression vs anxiety . . . 10
Total vs depression . . . 11
Total vs anxiety . . . 12
Point estimates . . . 13
Full cohort . . . 13
By HIV status . . . 14
95% CI of the point estimates . . . 16
Whole cohort . . . 16
By HIV status . . . 16
95% CI of the difference in mean . . . 16

Categorical data **18**
Sex (self-identified) . . . 18
Point estimates . . . 18
Full cohort . . . 18
By HIV status . . . 19
95% confidence intervals for the point estimates . . . 21
Full cohort . . . 21
By HIV status . . . 21
95% CI of the difference in proportions . . . 22
School grade . . . 23

Point estimates . . . 23
Full cohort . . . 23
By HIV status . . . 25
95% confidence intervals for the point estimates . . . 26
Full cohort . . . 26
By HIV status . . . 27
95% confidence interval of the difference in proportions . . . 29
Employment . . . 30
Point estimates . . . 30
Full cohort . . . 30
By HIV status . . . 33
95% confidence intervals for the point estimates . . . 34
Full cohort . . . 34
By HIV status . . . 36
95% confidence interval of the difference in proportions . . . 37

Session information **39**

This script generates summaries of key demographic information for the full cohort, with and without conditioning on HIV status.

We present the data in tabular and graphical format, and calculate the precision of the estimates using bootstrap 95% confidence intervals.

To describe any differences between the HIV+ and HIV- groups, we have calculated 95% confidence intervals of the difference in mean/proportion.

Import data

```
# Import data
general <- read_rds('data-cleaned/general_info.rds') %>%
    select(PID, age, sex, educational_level, employment)

mental_health <- read_rds('data-cleaned/hscl.rds') %>%
    select(PID, anxiety_score, depression_score, total_score)

# Join to core_info
data <- read_rds('data-cleaned/hiv_test.rds') %>%
    select(PID, test_result, CD4_count) %>%
    left_join(general) %>%
    left_join(mental_health)
```

Quick look

```
# Dataframe dimensions
dim(data)
```

```
## [1] 539  10
```

```
# Column names
names(data)
```

```
##  [1] "PID"               "test_result"       "CD4_count"
##  [4] "age"               "sex"               "educational_level"
##  [7] "employment"        "anxiety_score"     "depression_score"
## [10] "total_score"
```

```
# Glimpse data
glimpse(data)
```

```
## Observations: 539
## Variables: 10
## $ PID               <chr> "001", "003", "004", "005", "006", "007", "0...
## $ test_result       <chr> "HIV negative", "HIV negative", "HIV negativ...
## $ CD4_count         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ age               <dbl> 35, 50, 38, 37, 30, 25, 39, 27, 23, 32, 36, ...
## $ sex               <chr> "male", "female", "male", "male", "male", "m...
## $ educational_level <ord> secondary school, no/primary school, seconda...
## $ employment        <chr> "unemployed", "disability grant", "employed"...
## $ anxiety_score     <dbl> 3.5, 1.2, 2.1, 1.0, 2.7, 1.5, 1.7, 2.0, 2.4,...
## $ depression_score  <dbl> 3.333333, 1.333333, 1.800000, 1.066667, 2.73...
## $ total_score       <dbl> 3.40, 1.28, 1.92, 1.04, 2.72, 1.64, 1.76, 2....
```

Check missingness

Full cohort

```
data %>%
    profile_missing() %>%
    mutate(pct_missing = round(100 * pct_missing)) %>%
    arrange(pct_missing)
```

```
## # A tibble: 10 x 3
##    feature           num_missing pct_missing
##    <fct>                   <int>       <dbl>
##  1 PID                         0           0
##  2 sex                         2           0
##  3 anxiety_score               2           0
##  4 test_result                 4           1
##  5 age                         3           1
##  6 employment                  3           1
##  7 depression_score            4           1
##  8 total_score                 5           1
##  9 educational_level          14           3
## 10 CD4_count                 474          88
```

```
# Remove rows with missing HIV test results (n = 4)
data %<>%
    filter(!is.na(test_result))
```

HIV-

```
data %>%
    select(-CD4_count) %>%
    filter(test_result == 'HIV negative') %>%
    profile_missing() %>%
    mutate(pct_missing = round(100 * pct_missing)) %>%
    arrange(pct_missing)
```

```
## # A tibble: 9 x 3
##   feature           num_missing pct_missing
##   <fct>                   <int>       <dbl>
## 1 PID                         0           0
## 2 test_result                 0           0
## 3 age                         2           0
## 4 sex                         1           0
## 5 anxiety_score               2           0
## 6 employment                  3           1
## 7 depression_score            4           1
## 8 total_score                 5           1
## 9 educational_level          14           3
```

HIV+

```
data %>%
    filter(test_result == 'HIV positive') %>%
    profile_missing() %>%
    mutate(pct_missing = round(100 * pct_missing)) %>%
    arrange(pct_missing)
```

```
## # A tibble: 10 x 3
##    feature           num_missing pct_missing
##    <fct>                   <int>       <dbl>
##  1 PID                         0           0
##  2 test_result                 0           0
##  3 educational_level           0           0
##  4 employment                  0           0
##  5 anxiety_score               0           0
##  6 depression_score            0           0
##  7 total_score                 0           0
##  8 age                         1           1
##  9 sex                         1           1
## 10 CD4_count                   5           7
```

Numeric data

Age

Point estimates

Full cohort

```r
# Tabular summary
data %>%
    select(age) %>%
    skim_to_wide() %>%
    select(-type, -hist) %>%
    kable(caption = 'Age (full cohort)') %>%
    kable_styling(latex_options = c('scale_down',
                                    'hold_position'))
```

Table 1: Age (full cohort)

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
age	3	532	535	34.3	11.11	15	26	32	41	72

```r
# Graphical summary
ggplot(data = data) +
    aes(x = 'Full cohort',
        y = age) +
    geom_boxplot(alpha = 0.7,
                 colour = '#000000',
                 fill = '#999999') +
    geom_point(position = position_jitter(height = 0),
               fill = '#FFFFFF',
               alpha = 0.7,
               stroke = 0.8,
               size = 2,
               shape = 21) +
    labs(subtitle = 'Age for the full cohort',
         y = 'Age (years)') +
    theme(axis.title.x = element_blank())
```

By HIV status

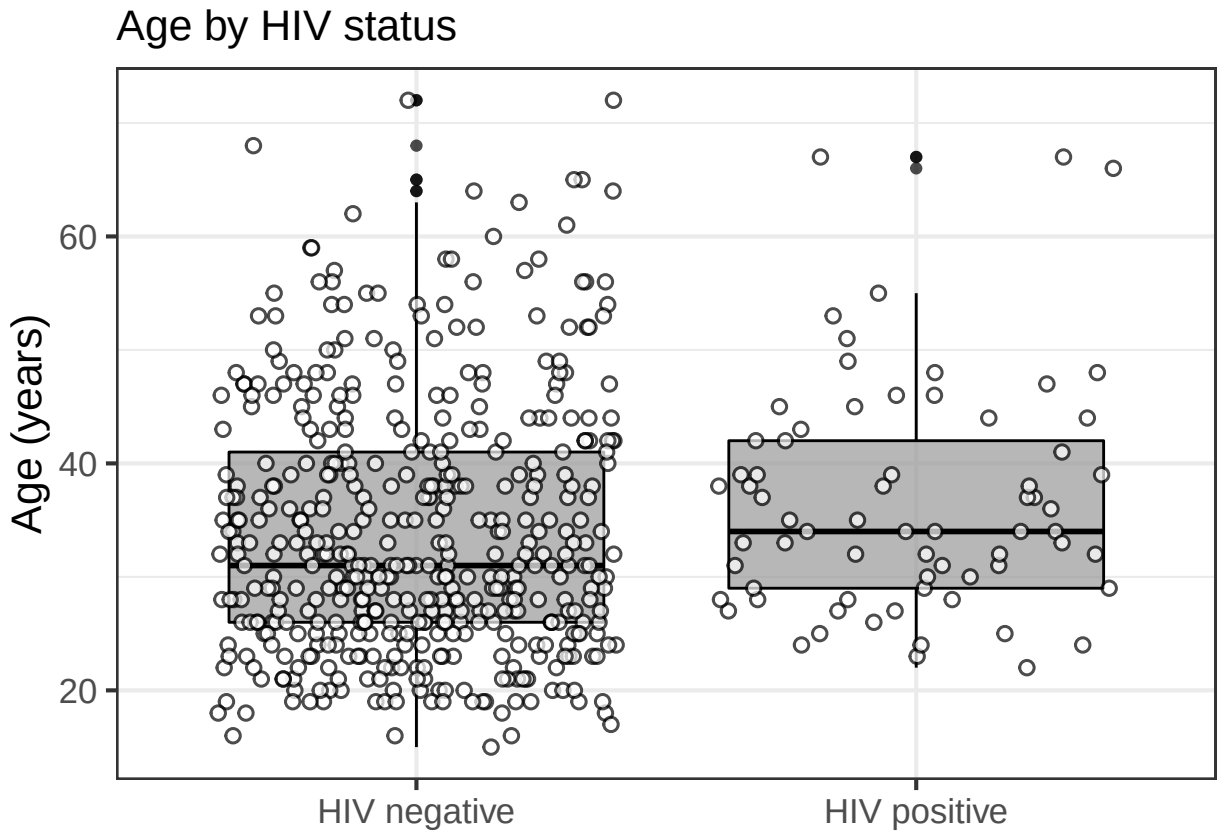
```
# Tabular summary
data %>%
    select(test_result, age) %>%
    group_by(test_result) %>%
    skim_to_wide() %>%
    select(-type, -hist) %>%
    kable(caption = 'Age (by HIV status)') %>%
    kable_styling(latex_options = c('scale_down',
                                    'hold_position'))
```

Table 2: Age (by HIV status)

test_result	variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
HIV negative	age	2	463	465	33.93	11.22	15	26	31	41	72
HIV positive	age	1	69	70	36.77	10.12	22	29	34	42	67

```
# Graphical summary
ggplot(data = data) +
    aes(x = test_result,
        y = age) +
    geom_boxplot(alpha = 0.7,
                 colour = '#000000',
                 fill = '#999999') +
    geom_point(position = position_jitter(height = 0),
               fill = '#FFFFFF',
```

```
            alpha = 0.7,
            stroke = 0.8,
            size = 2,
            shape = 21) +
  labs(subtitle = 'Age by HIV status',
       y = 'Age (years)') +
  theme(axis.title.x = element_blank(),
        legend.position = 'none')
```



95% CI of the point estimates

Whole cohort

```
groupwiseMean(age ~ 1,
              data = data,
              percent = TRUE) %>%
    select(-.id, -starts_with('Trad'))
```

```
##     n Mean Conf.level Percentile.lower Percentile.upper
## 1 532 34.3       0.95             33.5             35.3
```

By HIV status

```
groupwiseMean(age ~ test_result,
              data = data,
              percent = TRUE) %>%
    select(-starts_with('Trad'))
```

```
##    test_result   n Mean Conf.level Percentile.lower Percentile.upper
## 1 HIV negative 463 33.9       0.95             32.9             35.0
## 2 HIV positive  69 36.8       0.95             34.5             39.2
```

95% CI of the difference in mean

```
# Boot function
func_tmp <- function(d, i){
    data <- d[i, ]
    data_hiv <- filter(data, test_result == 'HIV positive')
    data_nohiv <- filter(data, test_result == 'HIV negative')
    mean_yes <- mean(data_hiv$age, na.rm = TRUE)
    mean_no <- mean(data_nohiv$age, na.rm = TRUE)
    mean_yes - mean_no
}

# Confidence interval of the difference in proportions (HIV+ minus HIV-)
set.seed(2019)
boot_tmp <- boot(data = data,
                 statistic = func_tmp,
                 R = 999,
                 stype = 'i')

bootci_tmp <- boot.ci(boot_tmp,
                      type = 'perc')

tibble_tmp <- tibble(`difference in mean` = round(boot_tmp$t0, 2),
                     `lower 95% CI` = round(bootci_tmp$percent[[4]], 2),
                     `upper 95% CI` = round(bootci_tmp$percent[[5]], 2))

tibble_tmp %>%
    kable(caption = 'Age (years) - 95% CI of the difference (HIV+ minus HIV-)')
```

\begin{table}[t]

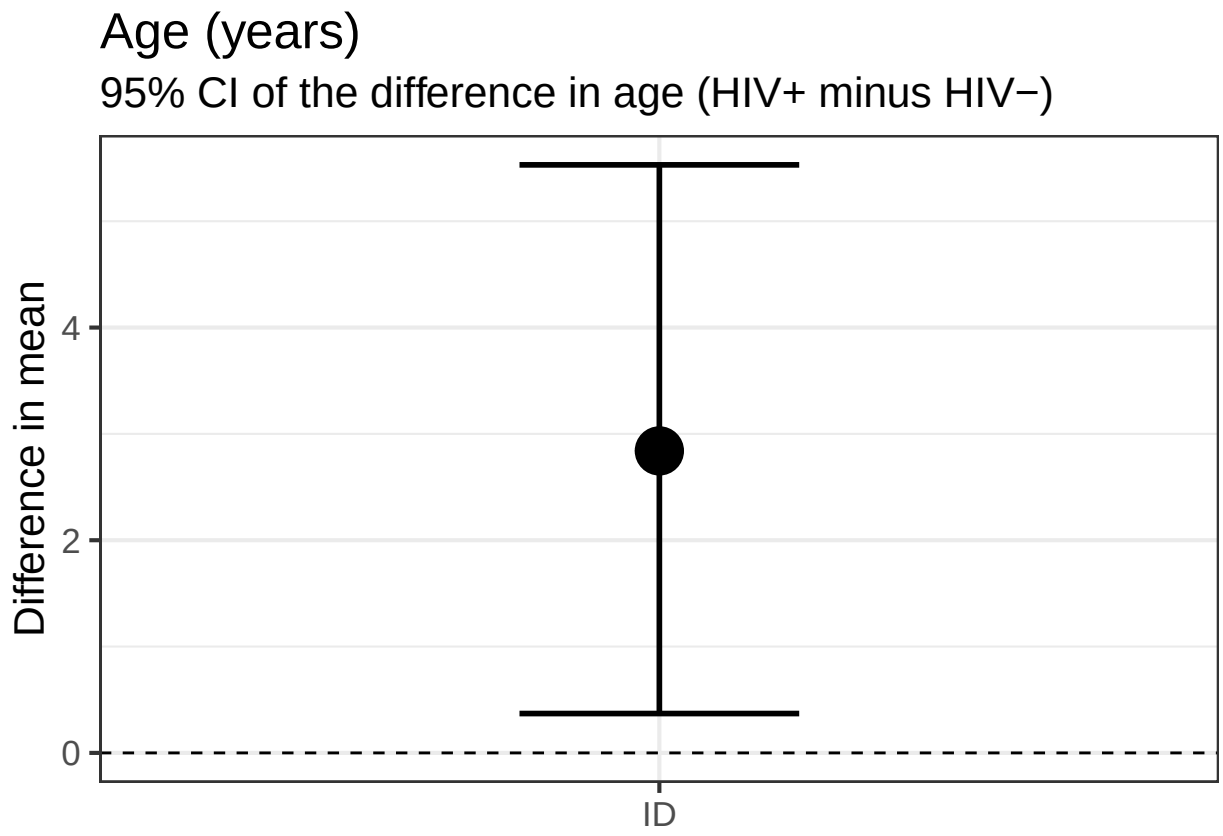
\caption{Age (years) - 95% CI of the difference (HIV+ minus HIV-)}

difference in mean	lower 95% CI	upper 95% CI
2.84	0.37	5.53

\end{table}

```
# Plot
ggplot(data = tibble_tmp) +
    aes(x = 'ID',
        y = `difference in mean`,
        ymin = `lower 95% CI`,
        ymax = `upper 95% CI`) +
    geom_point(size = 8) +
    geom_errorbar(size = 1,
                  width = 0.3) +
    geom_hline(yintercept = 0,
               linetype = 2) +
    labs(title = 'Age (years)',
         subtitle = '95% CI of the difference in age (HIV+ minus HIV-)',
```

```
      y = 'Difference in mean') +
  theme(axis.title.x = element_blank())
```



CD4 T-cell count

Only for participants that tested positive for HIV.

Point estimates

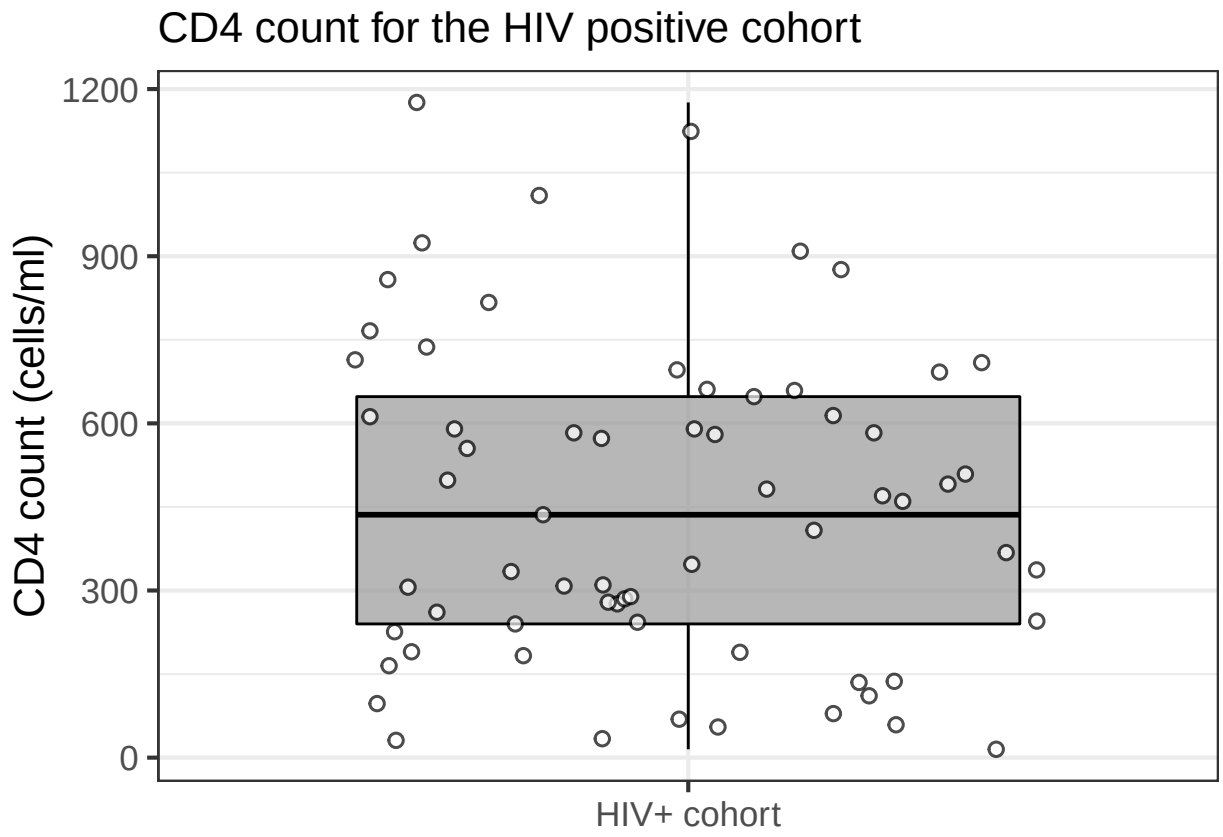
```
# Tabular summary
data %>%
    filter(test_result != 'HIV negative') %>%
    select(CD4_count) %>%
    skim_to_wide() %>%
    select(-type, -hist) %>%
    kable(caption = 'Age (full cohort)') %>%
    kable_styling(latex_options = c('scale_down',
                                    'hold_position'))
```

Table 3: Age (full cohort)

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
CD4_count	5	65	70	449.42	286.28	15	240	436	648	1176

```
# Graphical summary
data %>%
```

```
    filter(test_result != 'HIV negative') %>%
    ggplot(data = .) +
    aes(x = 'HIV+ cohort',
        y = CD4_count) +
    geom_boxplot(alpha = 0.7,
                 colour = '#000000',
                 fill = '#999999') +
    geom_point(position = position_jitter(height = 0),
               fill = '#FFFFFF',
               alpha = 0.7,
               stroke = 0.8,
               size = 2,
               shape = 21) +
    labs(subtitle = 'CD4 count for the HIV positive cohort',
         y = 'CD4 count (cells/ml)') +
    theme(axis.title.x = element_blank())
```



95% CI of the point estimates

```
cd4 <- data %>%
    filter(test_result != 'HIV negative') %>%
    filter(!is.na(CD4_count))

groupwiseMedian(CD4_count ~ 1,
                data = cd4,
                bca = FALSE,
```

```
                percent = TRUE) %>%
    select(-.id, -starts_with('Trad'))
```

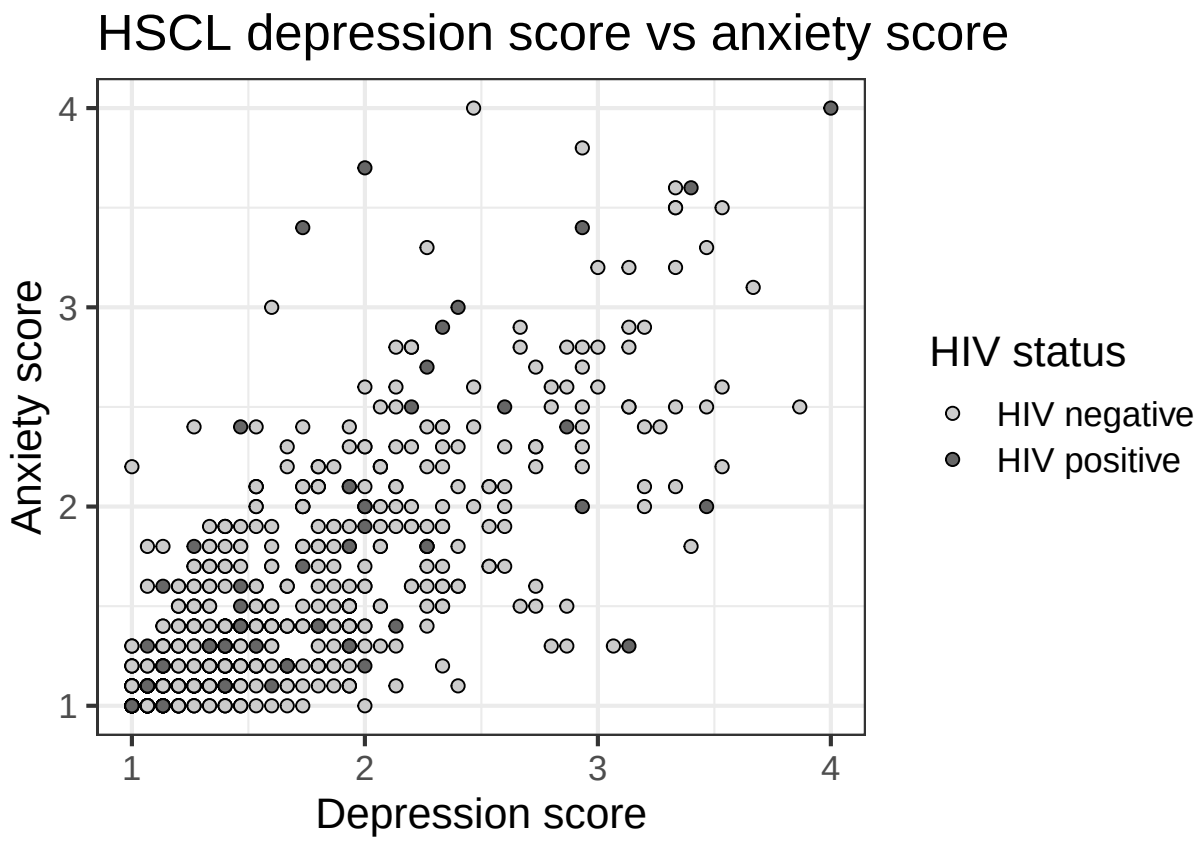
```
##    n Median Conf.level Percentile.lower Percentile.upper
## 1 65    436       0.95              306              573
```

Hopkins Symptom Checklist 25

Check whether anxiety and depression subscales are correlated with each other and with the total score.

Depression vs anxiety

```
# Scatterplot
ggplot(data = data) +
    aes(x = depression_score,
        y = anxiety_score) +
    geom_point(aes(fill = test_result),
               shape = 21,
               size = 2) +
    labs(title = 'HSCL depression score vs anxiety score',
         x = 'Depression score',
         y = 'Anxiety score') +
    scale_fill_manual(values = pal,
                      name = 'HIV status')
```

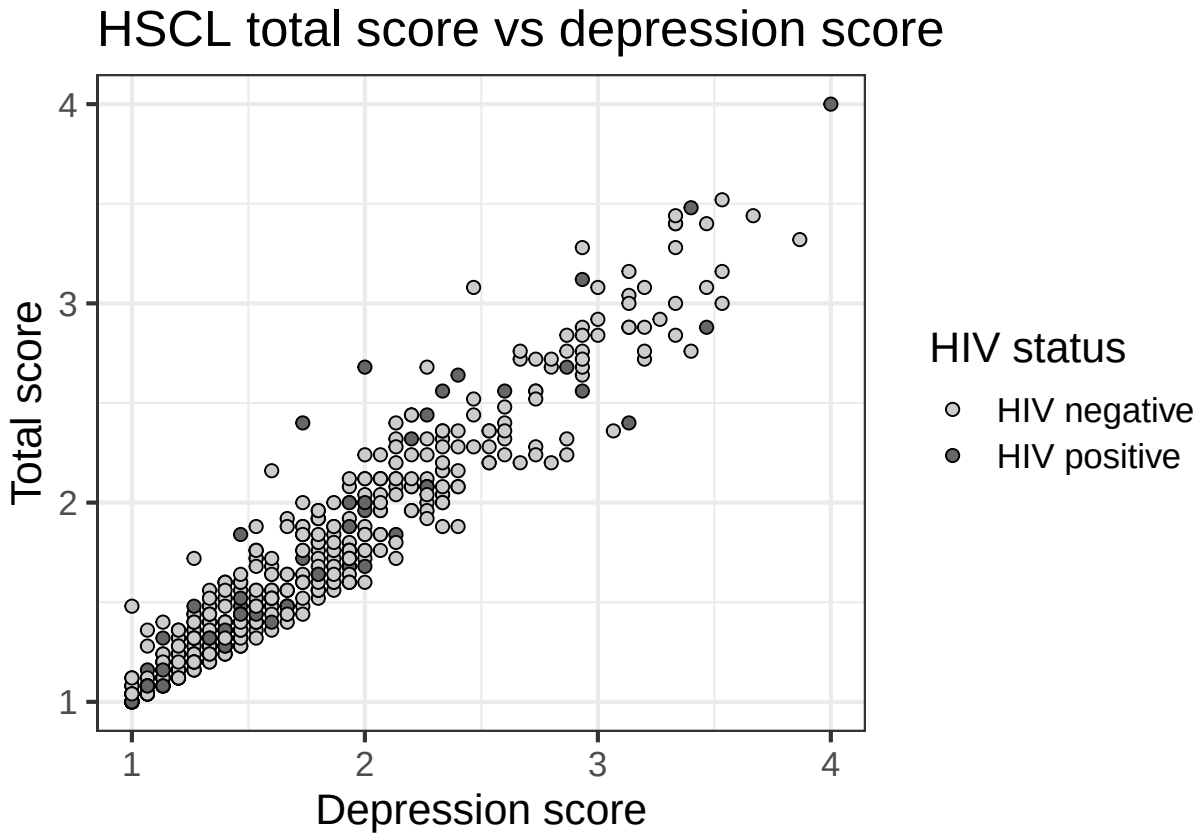


```r
# Correlation
with(data, cor.test(depression_score, anxiety_score))
```

```
##
##  Pearson's product-moment correlation
##
## data:  depression_score and anxiety_score
## t = 25.239, df = 528, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6982537 0.7757623
## sample estimates:
##       cor
## 0.7394487
```

Total vs depression

```r
# Scatterplot
ggplot(data = data) +
    aes(x = depression_score,
        y = total_score) +
    geom_point(aes(fill = test_result),
               shape = 21,
               size = 2) +
    labs(title = 'HSCL total score vs depression score',
         x = 'Depression score',
         y = 'Total score') +
    scale_fill_manual(values = pal,
                      name = 'HIV status')
```



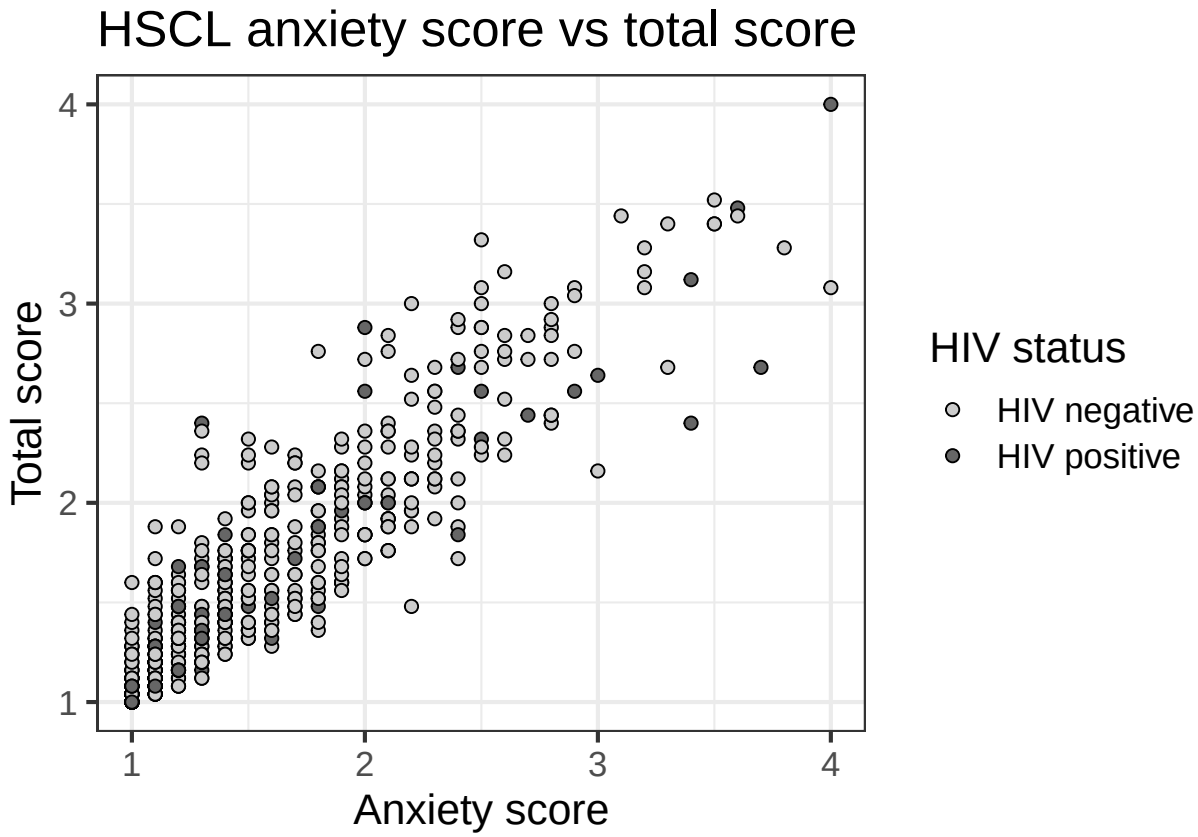

```
# Correlation
with(data, cor.test(total_score, depression_score))
```

```
##
##  Pearson's product-moment correlation
##
## data:  total_score and depression_score
## t = 79.476, df = 528, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9534984 0.9667276
## sample estimates:
##       cor
## 0.9606542
```

Total vs anxiety

```
# Scatterplot
ggplot(data = data) +
    aes(x = anxiety_score,
        y = total_score) +
    geom_point(aes(fill = test_result),
               shape = 21,
               size = 2) +
    labs(title = 'HSCL anxiety score vs total score',
         x = 'Anxiety score',
         y = 'Total score') +
```

```
    scale_fill_manual(values = pal,
                      name = 'HIV status')
```




```
# Correlation
with(data, cor.test(total_score, anxiety_score))
```

```
##
##  Pearson's product-moment correlation
##
## data:  total_score and anxiety_score
## t = 46.719, df = 528, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8793747 0.9127496
## sample estimates:
##       cor
## 0.8973375
```

Summary: Use the total HSCL score.

Point estimates

Full cohort

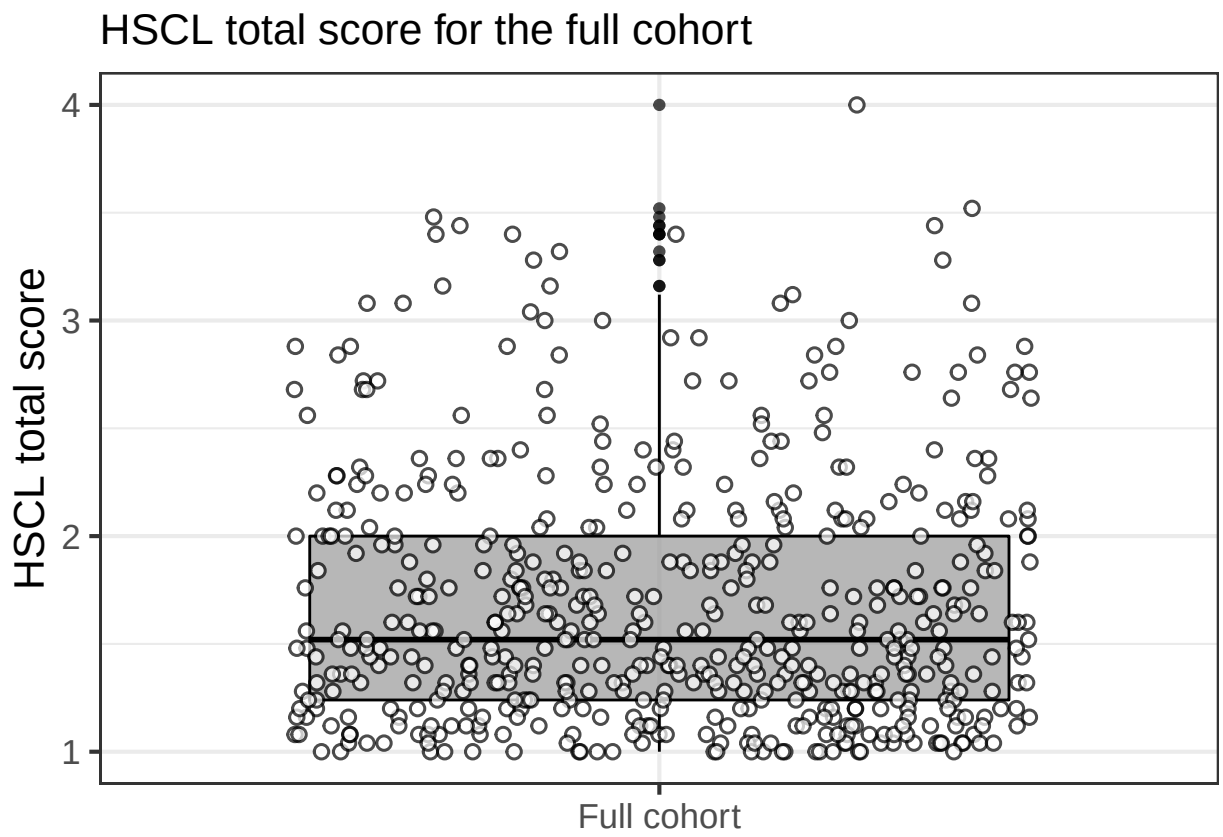
```
# Tabular summary
data %>%
    select(total_score) %>%
    skim_to_wide() %>%
    select(-type, -hist) %>%
```

```
    kable(caption = 'HSCL total score (full cohort)') %>%
    kable_styling(latex_options = c('scale_down',
                                    'hold_position'))
```

Table 4: HSCL total score (full cohort)

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
total_score	5	530	535	1.68	0.58	1	1.24	1.52	2	4

```
# Graphical summary
ggplot(data = data) +
    aes(x = 'Full cohort',
        y = total_score) +
    geom_boxplot(alpha = 0.7,
                 colour = '#000000',
                 fill = '#999999') +
    geom_point(position = position_jitter(height = 0),
               fill = '#FFFFFF',
               alpha = 0.7,
               stroke = 0.8,
               size = 2,
               shape = 21) +
    labs(subtitle = 'HSCL total score for the full cohort',
         y = 'HSCL total score') +
    theme(axis.title.x = element_blank())
```



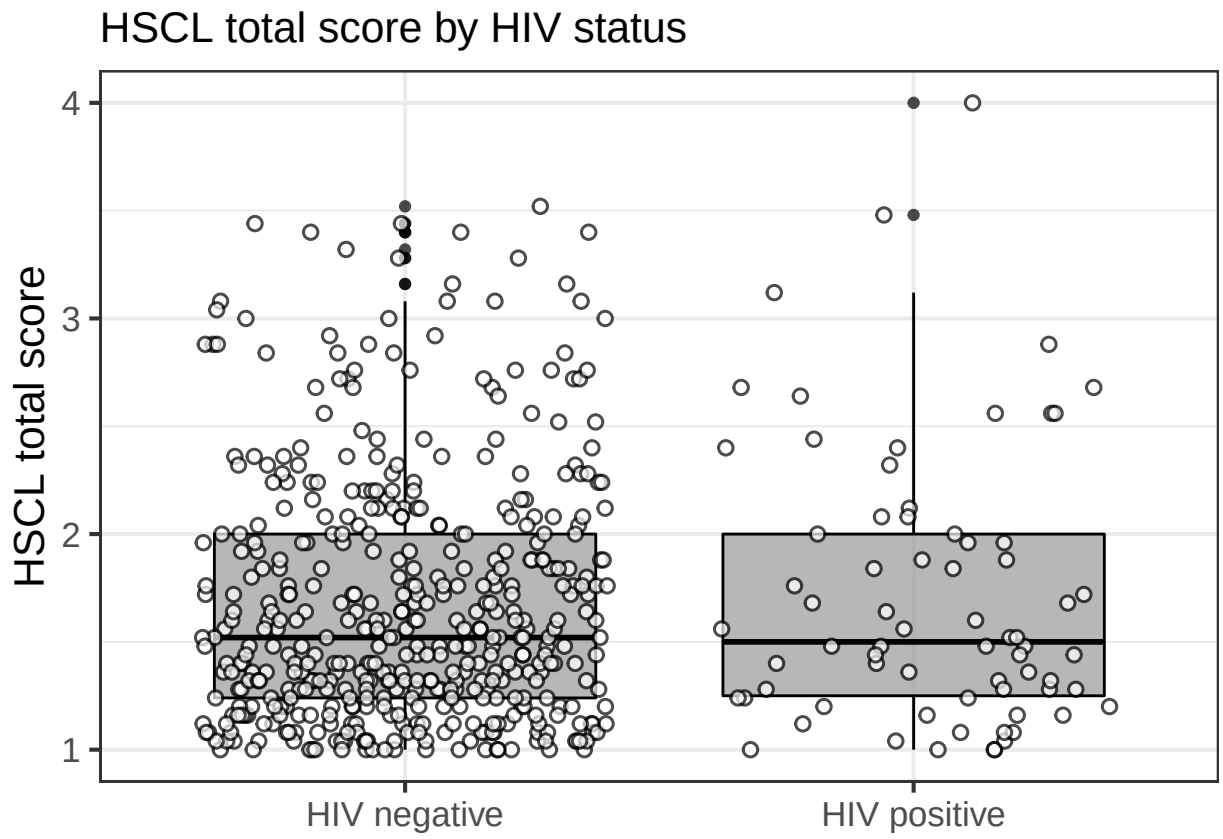
By HIV status

```r
# Tabular summary
data %>%
    select(test_result, total_score) %>%
    group_by(test_result) %>%
    skim_to_wide() %>%
    select(-type, -hist) %>%
    kable(caption = 'HSCL total score (by HIV status)') %>%
    kable_styling(latex_options = c('scale_down',
                                    'hold_position'))
```

Table 5: HSCL total score (by HIV status)

test_result	variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
HIV negative	total_score	5	460	465	1.68	0.57	1	1.24	1.52	2	3.52
HIV positive	total_score	0	70	70	1.72	0.64	1	1.25	1.5	2	4

```r
# Graphical summary
ggplot(data = data) +
    aes(x = test_result,
        y = total_score) +
    geom_boxplot(alpha = 0.7,
                 colour = '#000000',
                 fill = '#999999') +
    geom_point(position = position_jitter(height = 0),
               fill = '#FFFFFF',
               alpha = 0.7,
               stroke = 0.8,
               size = 2,
               shape = 21) +
    labs(subtitle = 'HSCL total score by HIV status',
         y = 'HSCL total score') +
    theme(axis.title.x = element_blank(),
          legend.position = 'none')
```



95% CI of the point estimates

Whole cohort

```
groupwiseMean(total_score ~ 1,
              data = data[!is.na(data$total_score), ],
              percent = TRUE) %>%
    select(-.id, -starts_with('Trad'))
```

```
##     n Mean Conf.level Percentile.lower Percentile.upper
## 1 530 1.68       0.95             1.64             1.73
```

By HIV status

```
groupwiseMean(total_score ~ test_result,
              data = data[!is.na(data$total_score), ],
              percent = TRUE) %>%
    select(-starts_with('Trad'))
```

```
##    test_result   n Mean Conf.level Percentile.lower Percentile.upper
## 1 HIV negative 460 1.68       0.95             1.63             1.73
## 2 HIV positive  70 1.72       0.95             1.57             1.87
```

95% CI of the difference in mean

```
# Boot function
func_tmp <- function(d, i){
```

```
    data <- d[i, ]
    data <- data[!is.na(data$total_score), ]
    data_hiv <- filter(data, test_result == 'HIV positive')
    data_nohiv <- filter(data, test_result == 'HIV negative')
    mean_yes <- mean(data_hiv$total_score, na.rm = TRUE)
    mean_no <- mean(data_nohiv$total_score, na.rm = TRUE)
    mean_yes - mean_no
}

# Confidence interval of the difference in proportions (HIV+ minus HIV-)
set.seed(2019)
boot_tmp <- boot(data = data,
                 statistic = func_tmp,
                 R = 999,
                 stype = 'i')

bootci_tmp <- boot.ci(boot_tmp,
                      type = 'perc') # BCa gave extreme order statistics

tibble_tmp <- tibble(`difference in mean` = round(boot_tmp$t0, 2),
                     `lower 95% CI` = round(bootci_tmp$percent[[4]], 2),
                     `upper 95% CI` = round(bootci_tmp$percent[[5]], 2))

tibble_tmp %>%
    kable(caption = 'HSCL total score - 95% CI of the difference (HIV+ minus HIV-)')
```

\begin{table}[t]

\caption{HSCL total score - 95% CI of the difference (HIV+ minus HIV-)}

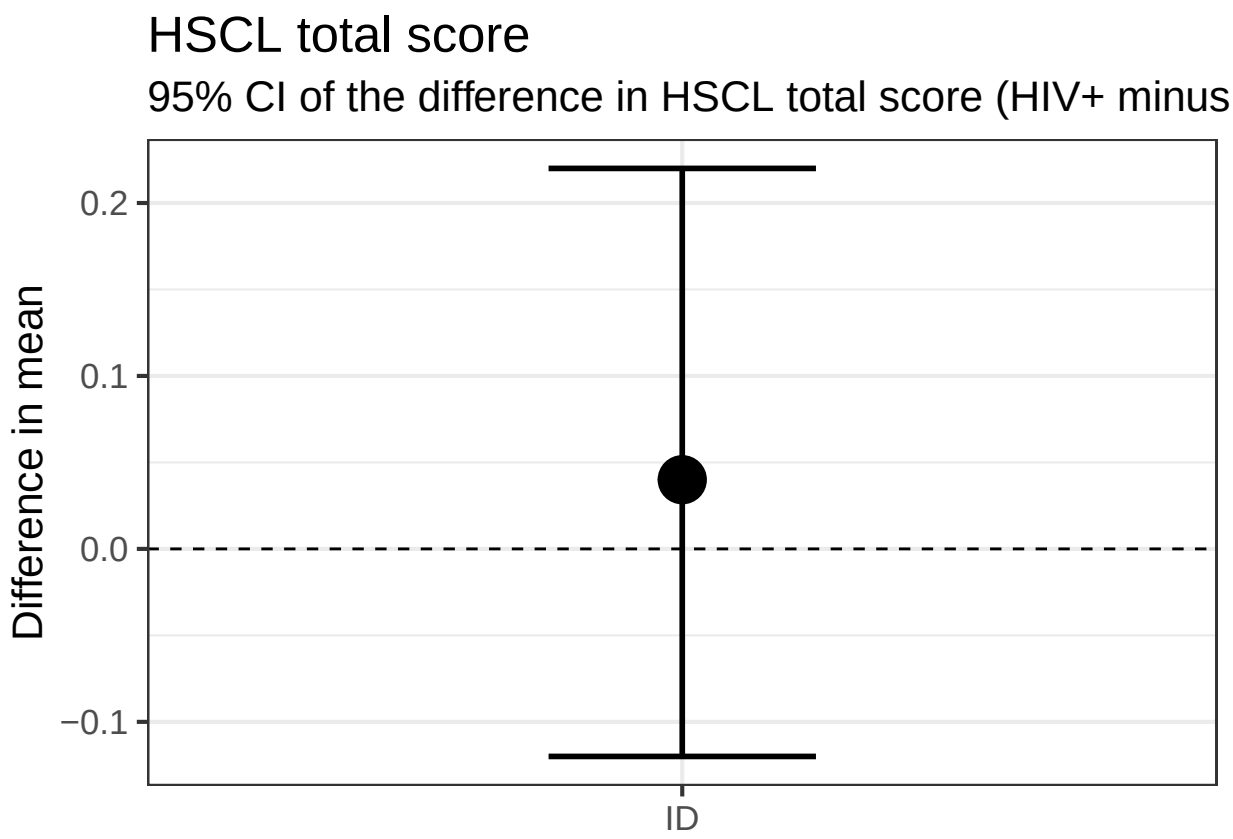
difference in mean	lower 95% CI	upper 95% CI
0.04	-0.12	0.22

\end{table}

```
# Plot
ggplot(data = tibble_tmp) +
    aes(x = 'ID',
        y = `difference in mean`,
        ymin = `lower 95% CI`,
        ymax = `upper 95% CI`) +
    geom_point(size = 8) +
    geom_errorbar(size = 1,
                  width = 0.3) +
    geom_hline(yintercept = 0,
               linetype = 2) +
    labs(title = 'HSCL total score',
         subtitle = '95% CI of the difference in HSCL total score (HIV+ minus HIV-)',
         y = 'Difference in mean') +
    theme(axis.title.x = element_blank())
```

Table 6: Sex - total cohort (point estimates)

sex	count	total	proportion
female	292	533	0.55
male	241	533	0.45



Categorical data

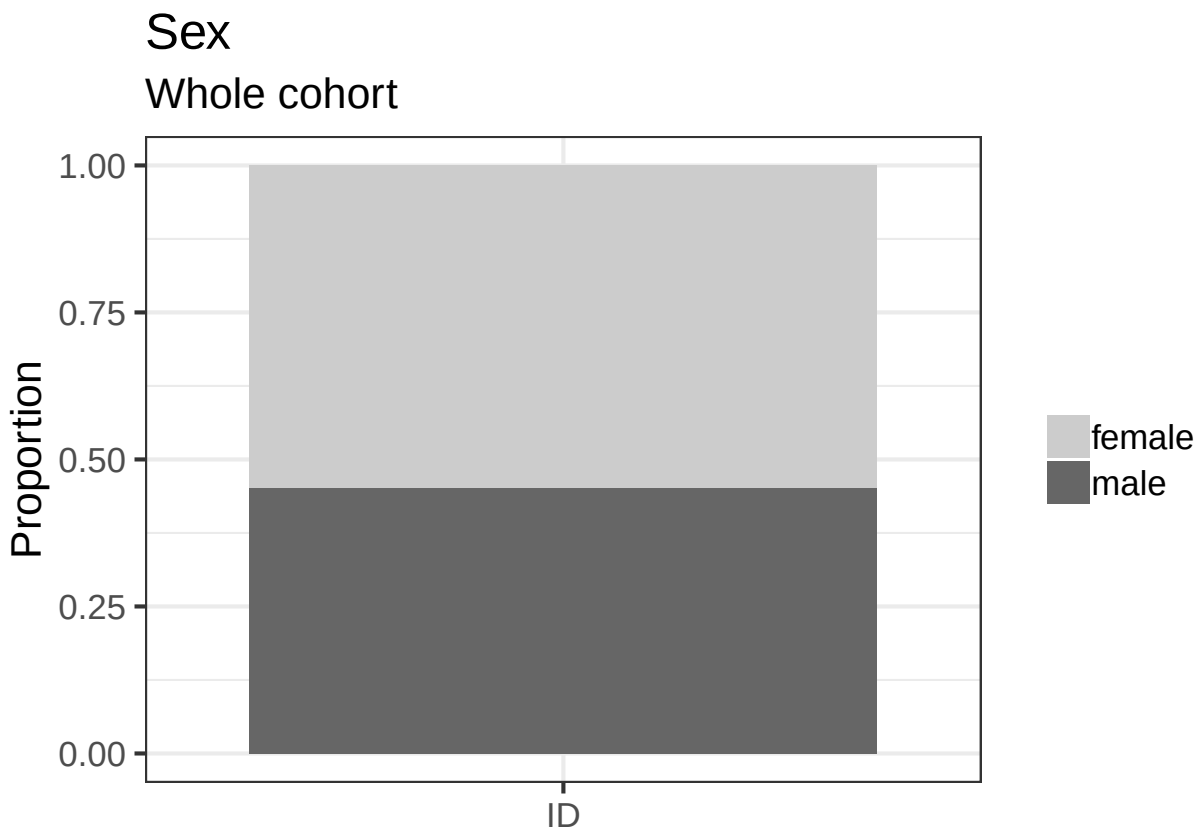
Sex (self-identified)

Point estimates

Full cohort

```
# Total cohort
data %>%
    filter(!is.na(sex)) %>%
    group_by(sex) %>%
    summarise(count = n()) %>%
    ungroup() %>%
    mutate(total = sum(count)) %>%
    mutate(proportion = round(count/total, 2)) %>%
    kable(caption = 'Sex - total cohort (point estimates)')
```

```r
## Plot
data %>%
    filter(!is.na(sex)) %>%
    ggplot(data = .) +
    aes(x = 'ID',
        fill = sex) +
    geom_bar(position = position_fill()) +
    scale_fill_manual(values = pal,
                      na.value = '#000000') +
    labs(title = 'Sex',
         subtitle = 'Whole cohort',
         y = 'Proportion') +
    theme(legend.title = element_blank(),
          axis.title.x = element_blank())
```



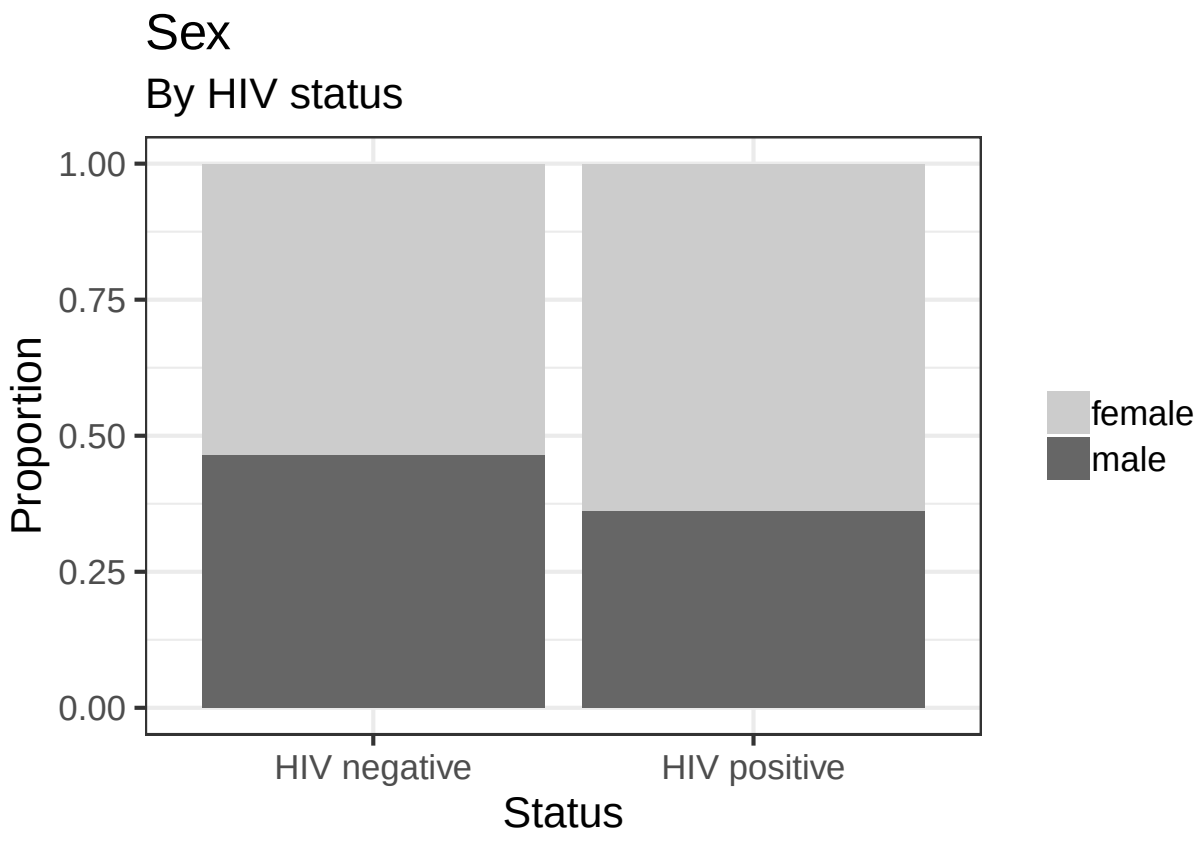
By HIV status

```r
# Count and proportion by HIV status
data %>%
    filter(!is.na(sex)) %>%
    filter(!is.na(test_result)) %>%
    group_by(test_result, sex) %>%
    summarise(count = n()) %>%
    group_by(test_result) %>%
    mutate(total = sum(count)) %>%
    mutate(proportion = round(count/total, 2)) %>%
    kable(caption = 'Sex - by HIV status (point estimates)')
```

Table 7: Sex - by HIV status (point estimates)

test_result	sex	count	total	proportion
HIV negative	female	248	464	0.53
HIV negative	male	216	464	0.47
HIV positive	female	44	69	0.64
HIV positive	male	25	69	0.36

```
## Plot
data %>%
    filter(!is.na(sex)) %>%
    filter(!is.na(test_result)) %>%
    ggplot(data = .) +
    aes(x = test_result,
        fill = sex) +
    geom_bar(position = position_fill()) +
    scale_fill_manual(values = pal,
                      na.value = '#000000') +
    labs(title = 'Sex',
         subtitle = 'By HIV status',
         y = 'Proportion',
         x = 'Status') +
    theme(legend.title = element_blank())
```



95% confidence intervals for the point estimates

Full cohort

```
# Boot functions
func_tmp <- function(d, i){
    data <- d[i, ]
    data <- data %>%
        filter(!is.na(sex))
    prop <- mean(data$sex == 'female')
    prop
}

# Whole cohort
set.seed(2019)
boot_tmp <- boot(data = data,
                 statistic = func_tmp,
                 R = 999,
                 stype = 'i')

bootci_tmp <- boot.ci(boot_tmp,
                      type = 'perc')

tibble(sex = 'female',
       proportion = round(boot_tmp$t0, 2),
       `lower 95% CI` = round(bootci_tmp$percent[[4]], 2),
       `upper 95% CI` = round(bootci_tmp$percent[[5]], 2)) %>%
    kable(caption = 'Sex - total cohort (95% CI)')
```

\begin{table}[t]

\caption{Sex - total cohort (95% CI)}

sex	proportion	lower 95% CI	upper 95% CI
female	0.55	0.51	0.59

\end{table}

By HIV status

```
# By HIV status (HIV- reported first)
set.seed(2019)
boot_tmp <- data %>%
    filter(!is.na(sex)) %>%
    filter(!is.na(test_result)) %>%
    group_by(test_result) %>%
    nest() %>%
    mutate(boot = map(.x = data,
                      ~ boot(data = .x,
                             statistic = func_tmp,
                             R = 999,
                             stype = 'i'))) %>%
    mutate(boot_ci = map(.x = boot,
                         ~ boot.ci(.x,
                                   type = 'perc')))
```

```
tibble(`status` = c('HIV negative', 'HIV positive'),
       sex = c('female', 'female'),
       proportion = c(round(boot_tmp$boot[[1]]$t0, 2),
                      round(boot_tmp$boot[[2]]$t0, 2)),
       `lower 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[4]], 2)),
       `upper 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[5]],2 ),
                          round(boot_tmp$boot_ci[[2]]$percent[[5]], 2))) %>%
    kable(caption = 'Sex - by HIV status (95% CI)')
```

\begin{table}[t]

\caption{Sex - by HIV status (95% CI)}

status	sex	proportion	lower 95% CI	upper 95% CI
HIV negative	female	0.53	0.49	0.58
HIV positive	female	0.64	0.52	0.75

\end{table}

95% CI of the difference in proportions

```
# Boot function
func_tmp <- function(d, i){
    data <- d[i, ]
    data <- data %>%
        filter(!is.na(sex)) %>%
        filter(!is.na(test_result))
    data_hiv <- filter(data, test_result == 'HIV positive')
    data_nohiv <- filter(data, test_result == 'HIV negative')
    prop_yes <- mean(data_hiv$sex == 'female')
    prop_no <- mean(data_nohiv$sex == 'female')
    prop_yes - prop_no
}

# Confidence interval of the difference in proportions (HIV+ minus HIV-)
set.seed(2019)
boot_tmp <- boot(data = data,
                 statistic = func_tmp,
                 R = 999,
                 stype = 'i')

bootci_tmp <- boot.ci(boot_tmp,
                      type = 'perc')

tibble_tmp <- tibble(`difference in proportion` = round(boot_tmp$t0, 2),
                     `lower 95% CI` = round(bootci_tmp$percent[[4]], 2),
                     `upper 95% CI` = round(bootci_tmp$percent[[5]], 2))

tibble_tmp %>%
    kable(caption = 'Sex - 95% CI of the difference (HIV+ minus HIV-)')
```

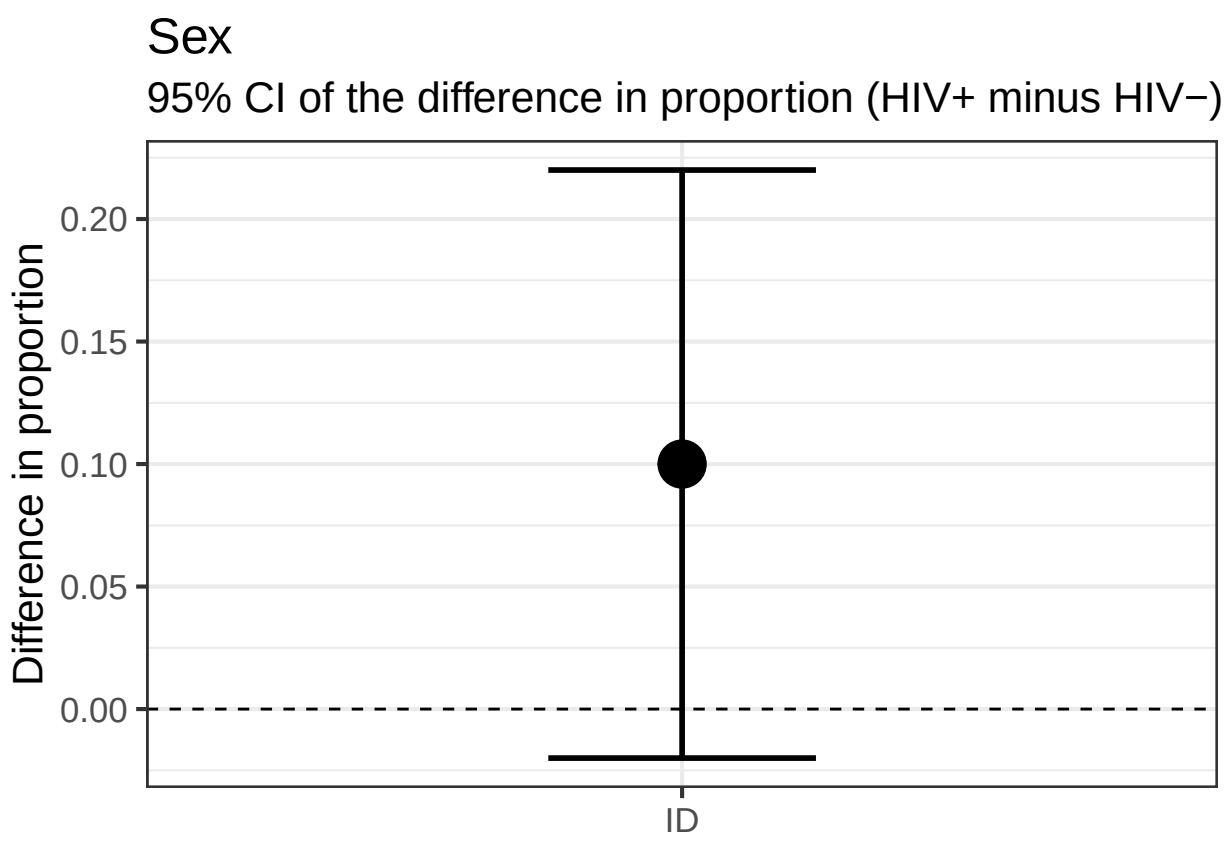
\begin{table}[t]

\caption{Sex - 95% CI of the difference (HIV+ minus HIV-)}

difference in proportion	lower 95% CI	upper 95% CI
0.1	-0.02	0.22

\end{table}

```
# Plot
ggplot(data = tibble_tmp) +
    aes(x = 'ID',
        y = `difference in proportion`,
        ymin = `lower 95% CI`,
        ymax = `upper 95% CI`) +
    geom_point(size = 8) +
    geom_errorbar(size = 1,
                  width = 0.3) +
    geom_hline(yintercept = 0,
               linetype = 2) +
    labs(title = 'Sex',
         subtitle = '95% CI of the difference in proportion (HIV+ minus HIV-)',
         y = 'Difference in proportion') +
    theme(axis.title.x = element_blank())
```



School grade

Point estimates

Full cohort

Table 8: Schooling - total cohort (point estimates)

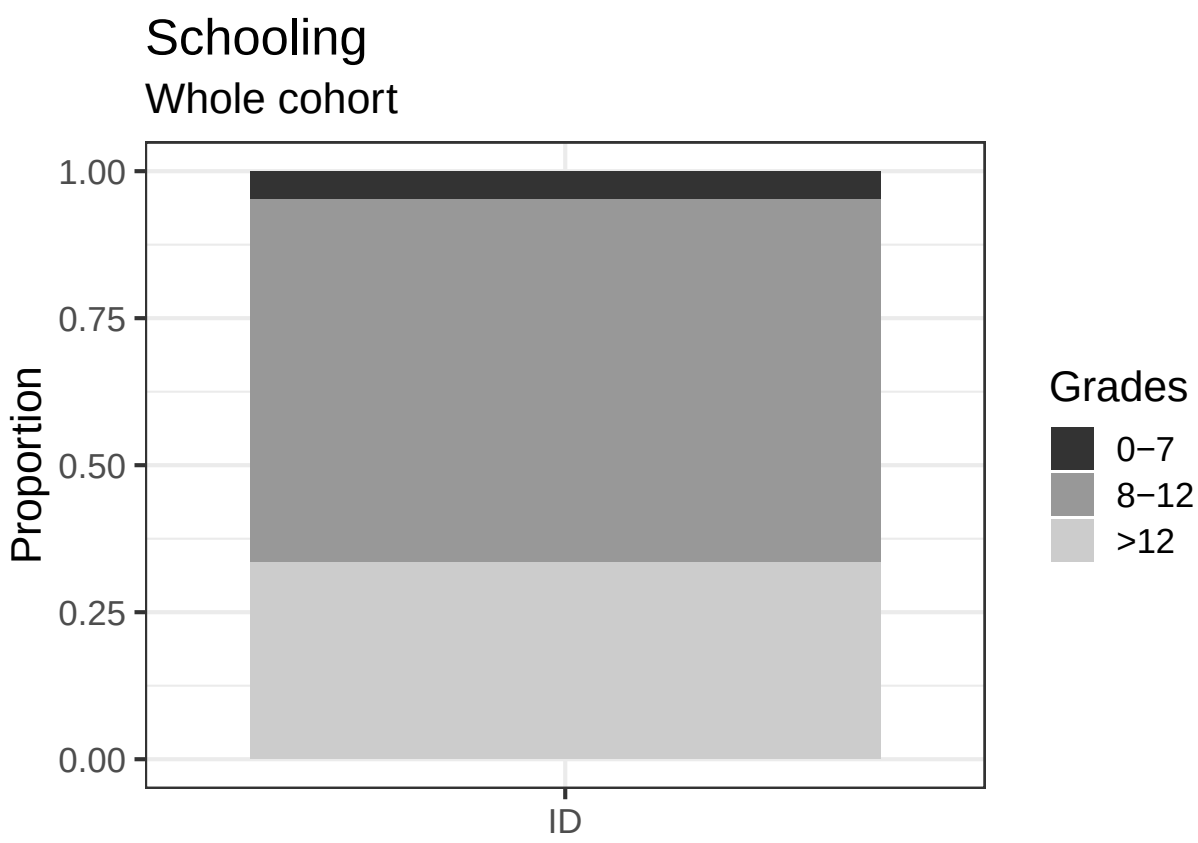
educational_level	count	total	proportion
no/primary school	24	521	0.05
secondary school	322	521	0.62
post-school qualification	175	521	0.34

```
# Total cohort
data %>%
    filter(!is.na(educational_level)) %>%
    group_by(educational_level) %>%
    summarise(count = n()) %>%
    ungroup() %>%
    mutate(total = sum(count)) %>%
    mutate(proportion = round(count/total, 2)) %>%
    kable(caption = 'Schooling - total cohort (point estimates)')

## Plot
data %>%
    filter(!is.na(educational_level)) %>%
    ggplot(data = .) +
    aes(x = 'ID',
        fill = educational_level) +
    geom_bar(position = position_fill()) +
    scale_fill_grey(name = 'Grades',
                    labels = c('0-7', '8-12', '>12')) +
    labs(title = 'Schooling',
         subtitle = 'Whole cohort',
         y = 'Proportion') +
    theme(axis.title.x = element_blank())
```

Table 9: Schooling - by HIV status (point estimates)

test_result	educational_level	count	total	proportion
HIV negative	no/primary school	14	451	0.03
HIV negative	secondary school	277	451	0.61
HIV negative	post-school qualification	160	451	0.35
HIV positive	no/primary school	10	70	0.14
HIV positive	secondary school	45	70	0.64
HIV positive	post-school qualification	15	70	0.21

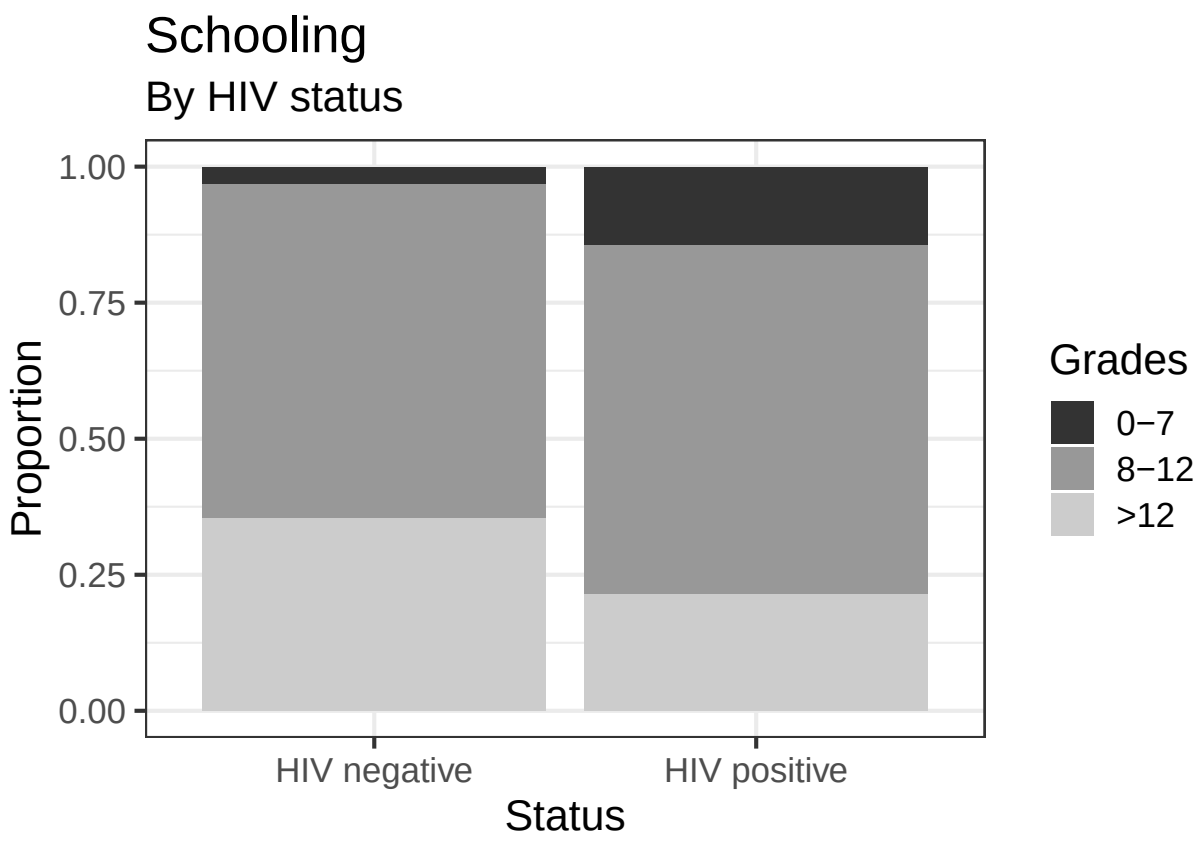


By HIV status

```
# Count and proportion by HIV status
data %>%
    filter(!is.na(educational_level)) %>%
    filter(!is.na(test_result)) %>%
    group_by(test_result, educational_level) %>%
    summarise(count = n()) %>%
    group_by(test_result) %>%
    mutate(total = sum(count)) %>%
    mutate(proportion = round(count/total, 2)) %>%
    kable(caption = 'Schooling - by HIV status (point estimates)')

## Plot
data %>%
    filter(!is.na(educational_level)) %>%
    filter(!is.na(test_result)) %>%
```

```
    ggplot(data = .) +
    aes(x = test_result,
        fill = educational_level) +
    geom_bar(position = position_fill()) +
    scale_fill_grey(name = 'Grades',
                    labels = c('0-7', '8-12', '>12')) +
    labs(title = 'Schooling',
         subtitle = 'By HIV status',
         y = 'Proportion',
         x = 'Status')
```



95% confidence intervals for the point estimates

Full cohort

```
school_tmp <- data %>%
    select(-CD4_count, -age, -sex, -employment,
           -anxiety_score, -depression_score, -total_score) %>%
    filter(!is.na(educational_level)) %>%
    mutate(dummy = row_number()) %>%
    spread(key = educational_level,
           value = dummy) %>%
    mutate_if(is.integer, ~ ifelse(!is.na(.),
                                   yes = 'yes',
                                   no = 'no')) %>%
    gather(key = grade,
           value = value,
```

```
           -PID, -test_result)

# Boot functions
func_tmp <- function(d, i){
    data <- d[i, ]
    data <- data %>%
        filter(!is.na(value))
    prop <- mean(data$value == 'yes')
    prop
}

# Whole cohort
set.seed(2019)
boot_tmp <- school_tmp %>%
    group_by(grade) %>%
    nest() %>%
    mutate(boot = map(.x = data,
                      ~ boot(data = .x,
                             statistic = func_tmp,
                             R = 999,
                             stype = 'i',
                             parallel = 'multicore',
                             ncpus = 7))) %>%
    mutate(boot_ci = map(.x = boot,
                         ~ boot.ci(.x,
                                   type = 'perc')))

tibble(grade = boot_tmp$grade,
       proportion = c(round(boot_tmp$boot[[1]]$t0, 2),
                      round(boot_tmp$boot[[2]]$t0, 2),
                      round(boot_tmp$boot[[3]]$t0, 2)),
       `lower 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[3]]$percent[[4]], 2)),
       `upper 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[3]]$percent[[5]], 2))) %>%
    kable(caption = 'Schooling - whole cohort (95% CI)')
```

\begin{table}[t]

\caption{Schooling - whole cohort (95% CI)}

grade	proportion	lower 95% CI	upper 95% CI
no/primary school	0.05	0.03	0.06
secondary school	0.62	0.58	0.66
post-school qualification	0.34	0.30	0.38

\end{table}

By HIV status

```
# Boot functions
func_tmp <- function(d, i){
    data <- d[i, ]
    data <- data %>%
```

```
        filter(!is.na(value))
    prop <- mean(data$value == 'yes')
    prop
}

# Whole cohort
set.seed(2019)
boot_tmp <- school_tmp %>%
    group_by(test_result, grade) %>%
    nest() %>%
    mutate(boot = map(.x = data,
                      ~ boot(data = .x,
                             statistic = func_tmp,
                             R = 999,
                             stype = 'i',
                             parallel = 'multicore',
                             ncpus = 7))) %>%
    mutate(boot_ci = map(.x = boot,
                         ~ boot.ci(.x,
                                   type = 'perc')))

tibble(test_result = boot_tmp$test_result,
       grade = boot_tmp$grade,
       proportion = c(round(boot_tmp$boot[[1]]$t0, 2),
                      round(boot_tmp$boot[[2]]$t0, 2),
                      round(boot_tmp$boot[[3]]$t0, 2),
                      round(boot_tmp$boot[[4]]$t0, 2),
                      round(boot_tmp$boot[[5]]$t0, 2),
                      round(boot_tmp$boot[[6]]$t0, 2)),
       `lower 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[3]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[4]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[5]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[6]]$percent[[4]], 2)),
       `upper 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[3]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[4]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[5]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[6]]$percent[[5]], 2))) %>%
    kable(caption = 'Schooling - by HIV status (95% CI)')
```

\begin{table}[t]

\caption{Schooling - by HIV status (95% CI)}

test_result	grade	proportion	lower 95% CI	upper 95% CI
HIV negative	no/primary school	0.03	0.02	0.05
HIV positive	no/primary school	0.14	0.06	0.23
HIV negative	secondary school	0.61	0.57	0.66
HIV positive	secondary school	0.64	0.53	0.74
HIV negative	post-school qualification	0.35	0.31	0.40
HIV positive	post-school qualification	0.21	0.11	0.31

\end{table}

95% confidence interval of the difference in proportions

```
# Boot function
func_tmp <- function(d, i){
    data <- d[i, ]
    data <- data %>%
        filter(!is.na(value)) %>%
        filter(!is.na(test_result))
    data_hiv <- filter(data, test_result == 'HIV positive')
    data_nohiv <- filter(data, test_result == 'HIV negative')
    prop_yes <- mean(data_hiv$value == 'yes')
    prop_no <- mean(data_nohiv$value == 'yes')
    prop_yes - prop_no
}

# Confidence interval of the difference in proportions (HIV+ minus HIV-)
set.seed(2019)
boot_tmp <- school_tmp %>%
    group_by(grade) %>%
    nest() %>%
    mutate(boot = map(.x = data,
                      ~ boot(data = .x,
                             statistic = func_tmp,
                             R = 999,
                             stype = 'i',
                             parallel = 'multicore',
                             ncpus = 7))) %>%
    mutate(boot_ci = map(.x = boot,
                         ~ boot.ci(.x,
                                   type = 'perc')))

tibble_tmp <- tibble(grade = boot_tmp$grade,
                     `difference in proportion` = c(round(boot_tmp$boot[[1]]$t0, 2),
                                                    round(boot_tmp$boot[[2]]$t0, 2),
                                                    round(boot_tmp$boot[[3]]$t0, 2)),
                     `lower 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[4]], 2),
                                        round(boot_tmp$boot_ci[[2]]$percent[[4]], 2),
                                        round(boot_tmp$boot_ci[[3]]$percent[[4]], 2)),
                     `upper 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[5]], 2),
                                        round(boot_tmp$boot_ci[[2]]$percent[[5]], 2),
                                        round(boot_tmp$boot_ci[[3]]$percent[[5]], 2))) %>%
    mutate(grade = factor(grade,
                          levels = c('no/primary school',
                                     'secondary school',
                                     'post-school qualification'),
                          labels = c('0-7', '8-12', '>12'),
                          ordered = TRUE))

tibble_tmp %>%
    kable(caption = 'Schooling - 95% CI of the difference (HIV+ minus HIV-)')
```

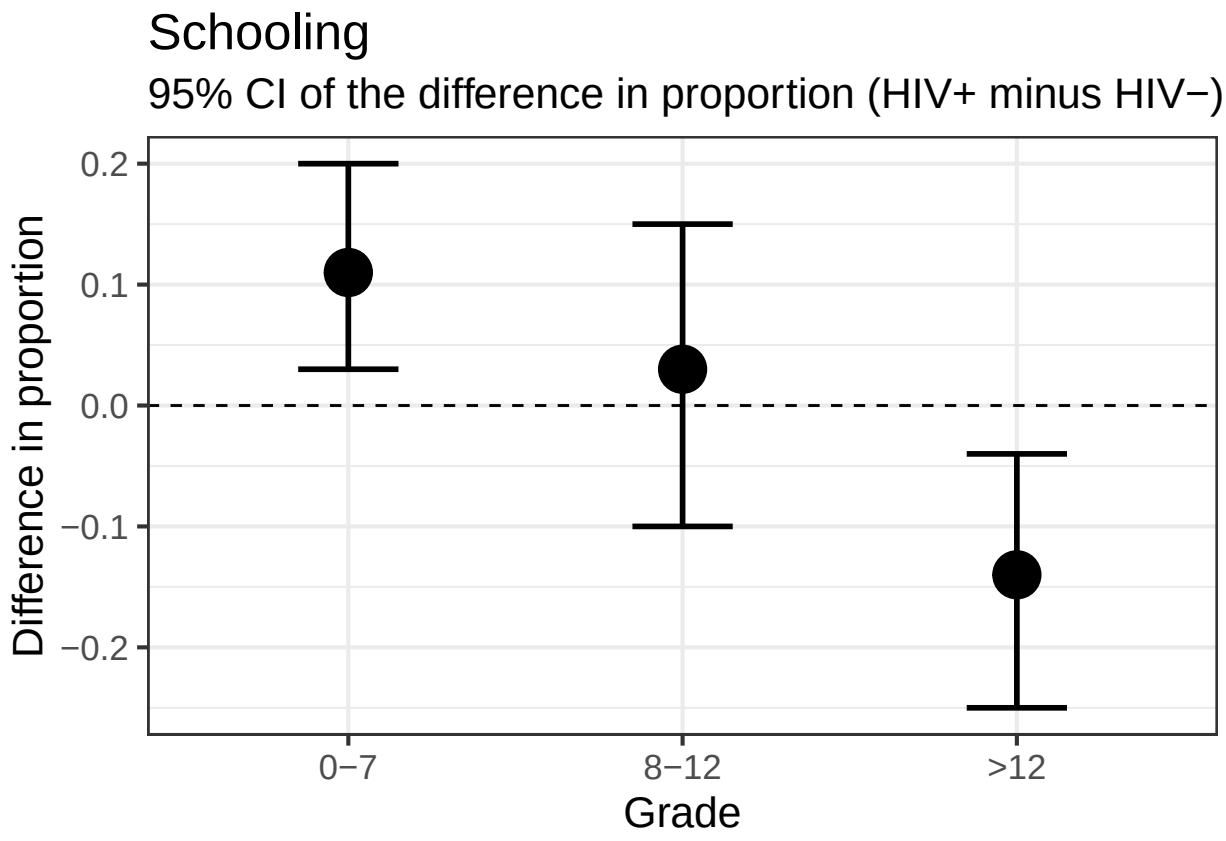
\begin{table}[t]

\caption{Schooling - 95% CI of the difference (HIV+ minus HIV-)}

grade	difference in proportion	lower 95% CI	upper 95% CI
0-7	0.11	0.03	0.20
8-12	0.03	-0.10	0.15
>12	-0.14	-0.25	-0.04

\end{table}

```
# Plot
ggplot(data = tibble_tmp) +
    aes(x = grade,
        y = `difference in proportion`,
        ymin = `lower 95% CI`,
        ymax = `upper 95% CI`) +
    geom_point(size = 8) +
    geom_errorbar(size = 1,
                  width = 0.3) +
    geom_hline(yintercept = 0,
               linetype = 2) +
    labs(title = 'Schooling',
         subtitle = '95% CI of the difference in proportion (HIV+ minus HIV-)',
         y = 'Difference in proportion',
         x = 'Grade')
```



Employment

Point estimates

Full cohort

Table 10: Employment - total cohort (point estimates)

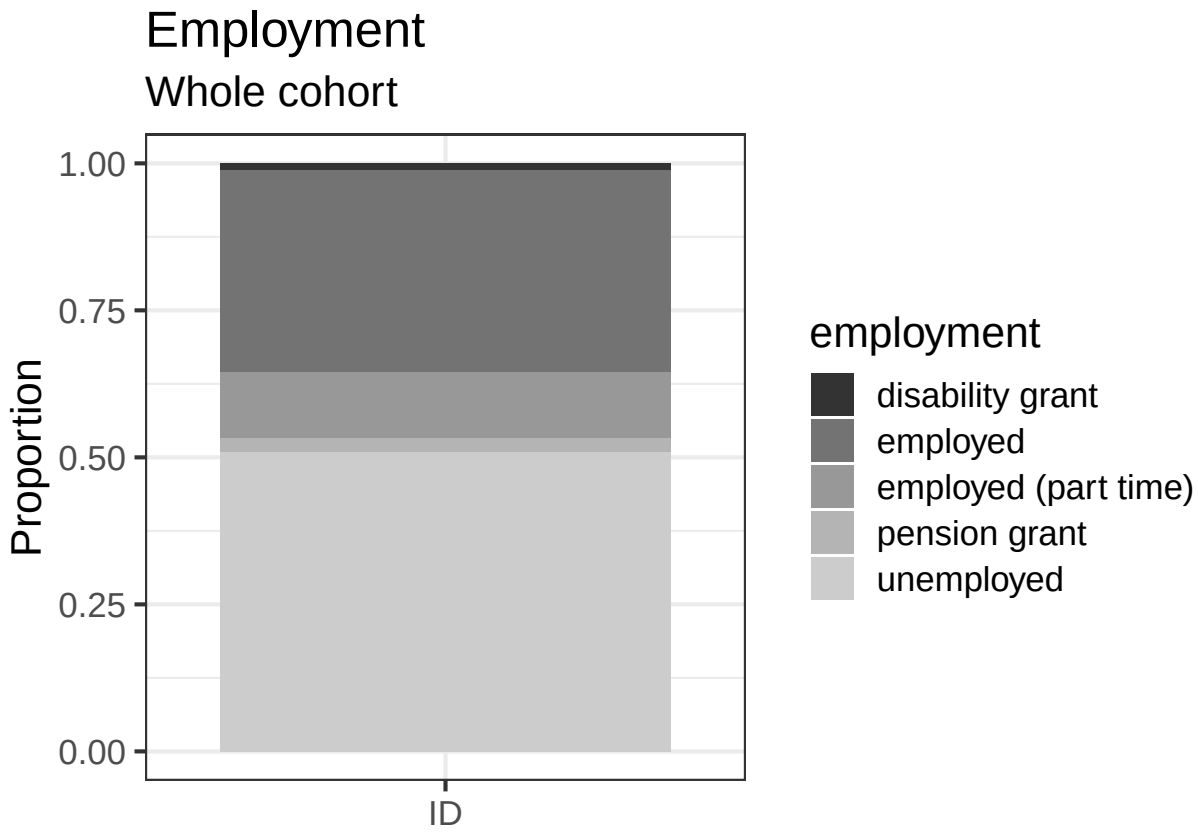
employment	count	total	proportion
disability grant	6	532	0.01
employed	182	532	0.34
employed (part time)	60	532	0.11
pension grant	13	532	0.02
unemployed	271	532	0.51

```
# Total cohort
data %>%
    filter(!is.na(employment)) %>%
    group_by(employment) %>%
    summarise(count = n()) %>%
    ungroup() %>%
    mutate(total = sum(count)) %>%
    mutate(proportion = round(count/total, 2)) %>%
    kable(caption = 'Employment - total cohort (point estimates)')

## Plot
data %>%
    filter(!is.na(employment)) %>%
    ggplot(data = .) +
    aes(x = 'ID',
        fill = employment) +
    geom_bar(position = position_fill()) +
    scale_fill_grey() +
    labs(title = 'Employment',
         subtitle = 'Whole cohort',
         y = 'Proportion') +
    theme(axis.title.x = element_blank())
```

Table 11: Employment (collapsed groups) - total cohort (point estimates)

employment	count	total	proportion
grant	19	532	0.04
employed	242	532	0.45
unemployed	271	532	0.51

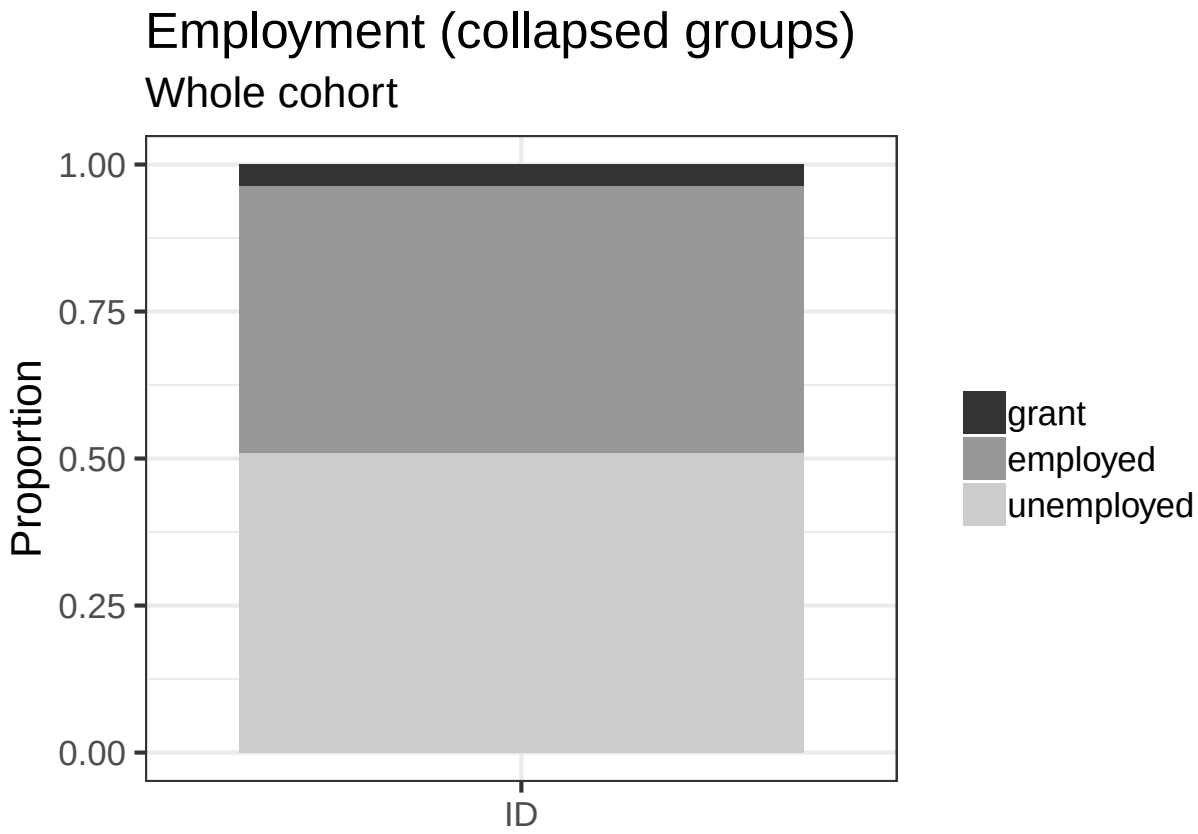



```
# Collapse the grants into one another and
# the same goes for the part-time/full-time employed categories
data %<>%
    mutate(employment = factor(employment),
           employment = fct_collapse(employment,
                                     employed = c('employed', 'employed (part time)'),
                                     grant = c('pension grant', 'disability grant')))

# Repeat analysis
# Total cohort
data %>%
    filter(!is.na(employment)) %>%
    group_by(employment) %>%
    summarise(count = n()) %>%
    ungroup() %>%
    mutate(total = sum(count)) %>%
    mutate(proportion = round(count/total, 2)) %>%
    kable(caption = 'Employment (collapsed groups) - total cohort (point estimates)')

## Plot
```

```
data %>%
    filter(!is.na(employment)) %>%
    ggplot(data = .) +
    aes(x = 'ID',
        fill = employment) +
    geom_bar(position = position_fill()) +
    scale_fill_grey() +
    labs(title = 'Employment (collapsed groups)',
         subtitle = 'Whole cohort',
         y = 'Proportion') +
    theme(axis.title.x = element_blank(),
          legend.title = element_blank())
```



By HIV status

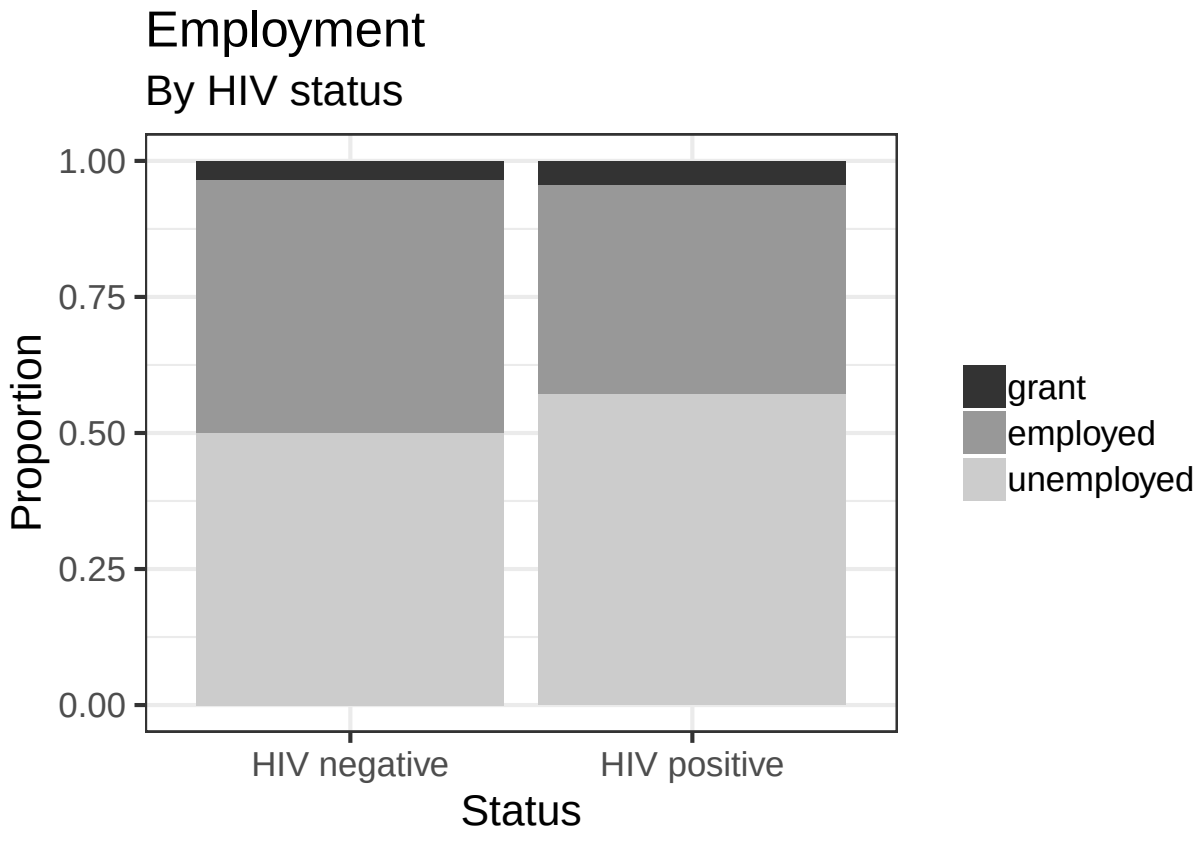
```
# Count and proportion by HIV status
data %>%
    filter(!is.na(employment)) %>%
    filter(!is.na(test_result)) %>%
    group_by(test_result, employment) %>%
    summarise(count = n()) %>%
    group_by(test_result) %>%
    mutate(total = sum(count)) %>%
    mutate(proportion = round(count/total, 2)) %>%
    kable(caption = 'Employment - by HIV status (point estimates)')

## Plot
data %>%
```

Table 12: Employment - by HIV status (point estimates)

test_result	employment	count	total	proportion
HIV negative	grant	16	462	0.03
HIV negative	employed	215	462	0.47
HIV negative	unemployed	231	462	0.50
HIV positive	grant	3	70	0.04
HIV positive	employed	27	70	0.39
HIV positive	unemployed	40	70	0.57

```
filter(!is.na(employment)) %>%
filter(!is.na(test_result)) %>%
ggplot(data = .) +
aes(x = test_result,
    fill = employment) +
geom_bar(position = position_fill()) +
scale_fill_grey() +
labs(title = 'Employment',
     subtitle = 'By HIV status',
     y = 'Proportion',
     x = 'Status') +
theme(legend.title = element_blank())
```



95% confidence intervals for the point estimates

Full cohort

```r
employment_tmp <- data %>%
    select(-CD4_count, -age, -sex, -educational_level,
           -anxiety_score, -depression_score, -total_score) %>%
    filter(!is.na(employment)) %>%
    mutate(dummy = row_number()) %>%
    spread(key = employment,
           value = dummy) %>%
    mutate_if(is.integer, ~ ifelse(!is.na(.),
                                   yes = 'yes',
                                   no = 'no')) %>%
    gather(key = employment,
           value = value,
           -PID, -test_result)

# Boot functions
func_tmp <- function(d, i){
    data <- d[i, ]
    data <- data %>%
        filter(!is.na(value))
    prop <- mean(data$value == 'yes')
    prop
}

# Whole cohort
set.seed(2019)
boot_tmp <- employment_tmp %>%
    group_by(employment) %>%
    nest() %>%
    mutate(boot = map(.x = data,
                      ~ boot(data = .x,
                             statistic = func_tmp,
                             R = 999,
                             stype = 'i',
                             parallel = 'multicore',
                             ncpus = 7))) %>%
    mutate(boot_ci = map(.x = boot,
                         ~ boot.ci(.x,
                                   type = 'perc')))

tibble(employment = boot_tmp$employment,
       proportion = c(round(boot_tmp$boot[[1]]$t0, 2),
                      round(boot_tmp$boot[[2]]$t0, 2),
                      round(boot_tmp$boot[[3]]$t0, 2)),
       `lower 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[3]]$percent[[4]], 2)),
       `upper 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[3]]$percent[[5]], 2))) %>%
    kable(caption = 'Employment - whole cohort (95% CI)')
```

\begin{table}[t]

\caption{Employment - whole cohort (95% CI)}

employment	proportion	lower 95% CI	upper 95% CI
grant	0.04	0.02	0.05
employed	0.45	0.41	0.50
unemployed	0.51	0.47	0.55

\end{table}

By HIV status

```
# Boot functions
func_tmp <- function(d, i){
    data <- d[i, ]
    data <- data %>%
        filter(!is.na(value))
    prop <- mean(data$value == 'yes')
    prop
}

# Whole cohort
set.seed(2019)
boot_tmp <- employment_tmp %>%
    group_by(test_result, employment) %>%
    nest() %>%
    mutate(boot = map(.x = data,
                      ~ boot(data = .x,
                             statistic = func_tmp,
                             R = 999,
                             stype = 'i',
                             parallel = 'multicore',
                             ncpus = 7))) %>%
    mutate(boot_ci = map(.x = boot,
                         ~ boot.ci(.x,
                                   type = 'perc')))

tibble(test_result = boot_tmp$test_result,
       grade = boot_tmp$employment,
       proportion = c(round(boot_tmp$boot[[1]]$t0, 2),
                      round(boot_tmp$boot[[2]]$t0, 2),
                      round(boot_tmp$boot[[3]]$t0, 2),
                      round(boot_tmp$boot[[4]]$t0, 2),
                      round(boot_tmp$boot[[5]]$t0, 2),
                      round(boot_tmp$boot[[6]]$t0, 2)),
       `lower 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[3]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[4]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[5]]$percent[[4]], 2),
                          round(boot_tmp$boot_ci[[6]]$percent[[4]], 2)),
       `upper 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[2]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[3]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[4]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[5]]$percent[[5]], 2),
                          round(boot_tmp$boot_ci[[6]]$percent[[5]], 2))) %>%
```

```
    kable(caption = 'Employment - by HIV status (95% CI)')
```

\begin{table}[t]

\caption{Employment - by HIV status (95% CI)}

test_result	grade	proportion	lower 95% CI	upper 95% CI
HIV negative	grant	0.03	0.02	0.05
HIV positive	grant	0.04	0.00	0.10
HIV negative	employed	0.47	0.42	0.51
HIV positive	employed	0.39	0.27	0.50
HIV negative	unemployed	0.50	0.45	0.54
HIV positive	unemployed	0.57	0.46	0.69

\end{table}

95% confidence interval of the difference in proportions

```
# Boot function
func_tmp <- function(d, i){
    data <- d[i, ]
    data <- data %>%
        filter(!is.na(value)) %>%
        filter(!is.na(test_result))
    data_hiv <- filter(data, test_result == 'HIV positive')
    data_nohiv <- filter(data, test_result == 'HIV negative')
    prop_yes <- mean(data_hiv$value == 'yes')
    prop_no <- mean(data_nohiv$value == 'yes')
    prop_yes - prop_no
}

# Confidence interval of the difference in proportions (HIV+ minus HIV-)
set.seed(2019)
boot_tmp <- employment_tmp %>%
    group_by(employment) %>%
    nest() %>%
    mutate(boot = map(.x = data,
                      ~ boot(data = .x,
                             statistic = func_tmp,
                             R = 999,
                             stype = 'i',
                             parallel = 'multicore',
                             ncpus = 7))) %>%
    mutate(boot_ci = map(.x = boot,
                         ~ boot.ci(.x,
                                   type = 'perc')))

tibble_tmp <- tibble(employment = boot_tmp$employment,
                     `difference in proportion` = c(round(boot_tmp$boot[[1]]$t0, 2),
                                                    round(boot_tmp$boot[[2]]$t0, 2),
                                                    round(boot_tmp$boot[[3]]$t0, 2)),
                     `lower 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[4]], 2),
                                        round(boot_tmp$boot_ci[[2]]$percent[[4]], 2),
                                        round(boot_tmp$boot_ci[[3]]$percent[[4]], 2)),
                     `upper 95% CI` = c(round(boot_tmp$boot_ci[[1]]$percent[[5]], 2),
```

```
                                        round(boot_tmp$boot_ci[[2]]$percent[[5]], 2),
                                        round(boot_tmp$boot_ci[[3]]$percent[[5]], 2)))

tibble_tmp %>%
    kable(caption = 'Employment - 95% CI of the difference (HIV+ minus HIV-)')
```

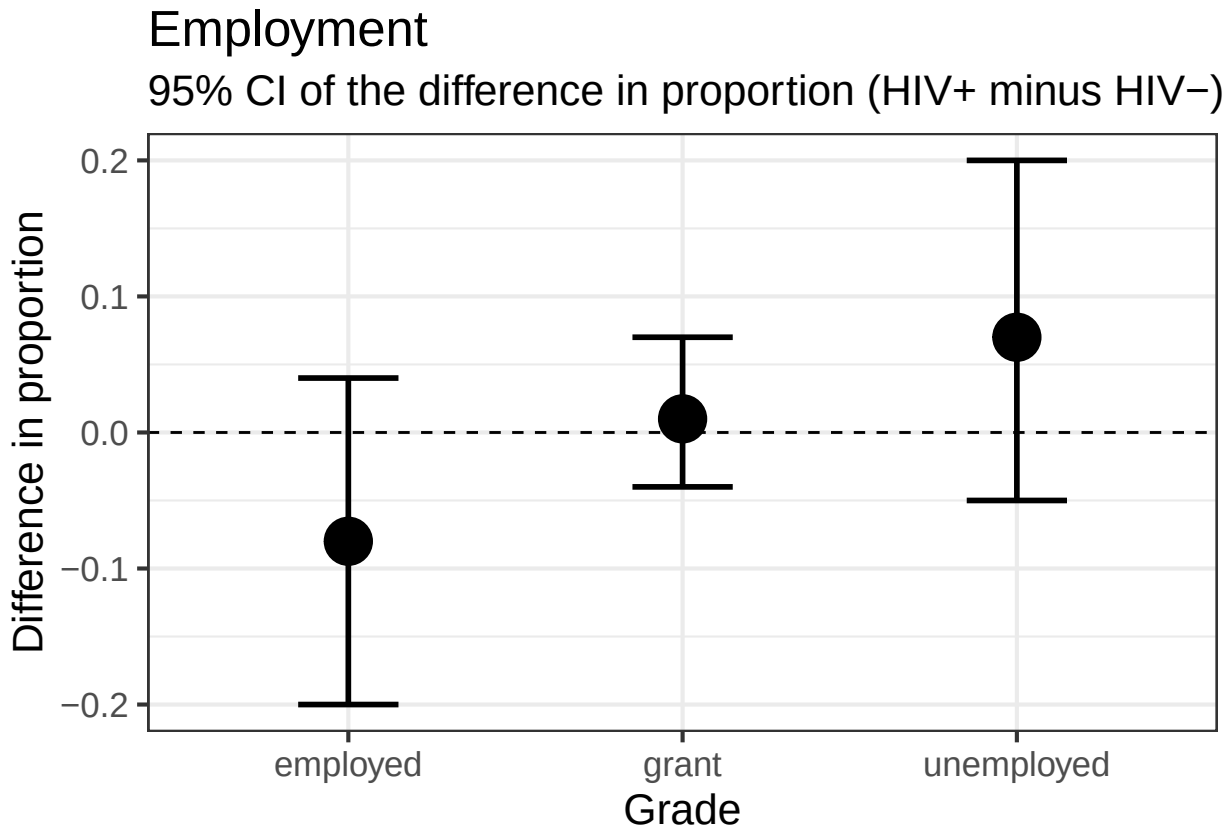
\begin{table}[t]

\caption{Employment - 95% CI of the difference (HIV+ minus HIV-)}

employment	difference in proportion	lower 95% CI	upper 95% CI
grant	0.01	-0.04	0.07
employed	-0.08	-0.20	0.04
unemployed	0.07	-0.05	0.20

\end{table}

```
# Plot
ggplot(data = tibble_tmp) +
    aes(x = employment,
        y = `difference in proportion`,
        ymin = `lower 95% CI`,
        ymax = `upper 95% CI`) +
    geom_point(size = 8) +
    geom_errorbar(size = 1,
                  width = 0.3) +
    geom_hline(yintercept = 0,
               linetype = 2) +
    labs(title = 'Employment',
         subtitle = '95% CI of the difference in proportion (HIV+ minus HIV-)',
         y = 'Difference in proportion',
         x = 'Grade')
```



Session information

```
sessionInfo()
```