Supplement 1

Descriptive statistics for the whole cohort

Peter Kamerman and Prinisha Pillay

17 March 2019

Contents

This script generates descriptive statistics for variables collected at baseline (visit day: 0, visit_number: 1) for the whole cohort, irrespective of whether they went on to develop sensory neuropathy (SN) or not.

The following data columns were not analysed:

- `pain` and `pain score`: related to SN only, therefore not relevant at baseline when everyone was free from SN.

- `*_record`: inconsistent patient records.

- `hba1c_percent` and `vitaminB12_pmol.l`: These data were only used to classify individuals as having diabetes or vitamin B12 deficiency when cleaning the data *(see:clean-data.R)*.

- `ID`, `visit_number`, `visit_day`, `hivsn_present`, `visit_months`: provide sorting and grouping information only.

Import data

```
data <- read_rds('data-cleaned/clean_data.rds') %>%
    # Filter for visit one
    filter(visit_number == 1) %>%
    # Remove columns that won't be analysed
```

```
    select(-starts_with('pain'), -ends_with('_record'),
           -visit_number, -visit_months,
           -hba1c_percent, -vitaminB12_pmol.l)
```

Inspect data

```
# Dimensions
dim(data)

## [1] 120  17

# Column names
names(data)

##  [1] "ID"                   "visit_day"
##  [3] "age_years"            "mass_kg"
##  [5] "height_m"             "sex"
##  [7] "hivsn_present"        "CD4_cell.ul"
##  [9] "viral_load_copies.ml" "consumes_alcohol"
## [11] "alcohol_units.week"   "TB_current"
## [13] "pyridoxine_prophylaxis" "rifafour_treatment"
## [15] "ARV_regimen"          "diabetic_hba1c"
## [17] "vitaminB12_deficiency"

# Head and tail
head(data)

## # A tibble: 6 x 17
##   ID    visit_day age_years mass_kg height_m sex   hivsn_present
##   <chr>     <int>     <dbl>   <dbl>    <dbl> <fct> <fct>
## 1 001           0        59    41.4     1.56 F     no
## 2 002           0        23    70.2     1.56 F     no
## 3 003           0        27    75       1.64 M     no
## 4 004           0        26    68.8     1.74 M     no
## 5 005           0        37   107       1.6  F     no
## 6 006           0        34    85.5     1.53 F     no
## # ... with 10 more variables: CD4_cell.ul <dbl>,
## #   viral_load_copies.ml <dbl>, consumes_alcohol <fct>,
## #   alcohol_units.week <int>, TB_current <fct>,
## #   pyridoxine_prophylaxis <fct>, rifafour_treatment <fct>,
## #   ARV_regimen <fct>, diabetic_hba1c <fct>, vitaminB12_deficiency <fct>

tail(data)

## # A tibble: 6 x 17
##   ID    visit_day age_years mass_kg height_m sex   hivsn_present
##   <chr>     <int>     <dbl>   <dbl>    <dbl> <fct> <fct>
## 1 115           0        29    55.1     1.66 M     no
## 2 116           0        30    93.7     1.55 F     no
## 3 117           0        30    58.2     1.6  F     no
## 4 118           0        30    61.2     1.64 F     no
## 5 119           0        22    62.7     1.63 F     no
## 6 120           0        58    71.2     1.74 M     no
## # ... with 10 more variables: CD4_cell.ul <dbl>,
## #   viral_load_copies.ml <dbl>, consumes_alcohol <fct>,
## #   alcohol_units.week <int>, TB_current <fct>,
```

```
## #    pyridoxine_prophylaxis <fct>, rifafour_treatment <fct>,
## #    ARV_regimen <fct>, diabetic_hba1c <fct>, vitaminB12_deficiency <fct>
```

```r
# Data structure
glimpse(data)
```

```
## Observations: 120
## Variables: 17
## $ ID                    <chr> "001", "002", "003", "004", "005", "006...
## $ visit_day             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ age_years             <dbl> 59, 23, 27, 26, 37, 34, 44, 34, 32, 29,...
## $ mass_kg               <dbl> 41.4, 70.2, 75.0, 68.8, 107.0, 85.5, 12...
## $ height_m              <dbl> 1.56, 1.56, 1.64, 1.74, 1.60, 1.53, 1.6...
## $ sex                   <fct> F, F, M, M, F, F, F, F, F, M, M, M, M, ...
## $ hivsn_present         <fct> no, no, no, no, no, no, no, no, no, no,...
## $ CD4_cell.ul           <dbl> 35, 285, 28, 270, 310, 247, 439, 311, 1...
## $ viral_load_copies.ml  <dbl> 6.103804, 5.041393, 5.181844, 2.484300,...
## $ consumes_alcohol      <fct> no, no, no, no, yes, no, no, no, no, no...
## $ alcohol_units.week    <int> 0, 0, 0, 0, 15, 0, 0, 0, 0, 0, 0, 6, 9,...
## $ TB_current            <fct> no, no, yes, no, no, no, no, no, no, no...
## $ pyridoxine_prophylaxis <fct> no, no, yes, no, no, no, no, no, no, no...
## $ rifafour_treatment    <fct> no, no, yes, no, no, no, no, no, no, no...
## $ ARV_regimen           <fct> TDF_FTC_EFV, TDF_FTC_EFV, TDF_FTC_EFV, ...
## $ diabetic_hba1c        <fct> no, no, no, no, no, no, no, no, no, no,...
## $ vitaminB12_deficiency <fct> no, no, no, no, no, no, no, no, no, no,...
```

Analyses

Age

```r
# Tabular summary
data %>%
    select(age_years) %>%
    skim()
```

```
## Skim summary statistics
##  n obs: 120
##  n variables: 1
##
## -- Variable type:numeric -------------------------------------------------------------------------
##   variable missing complete    n  mean   sd p0  p25 p50 p75 p100     hist
##  age_years       0      120  120 37.77 9.36 21 30.75  37  44   59
```
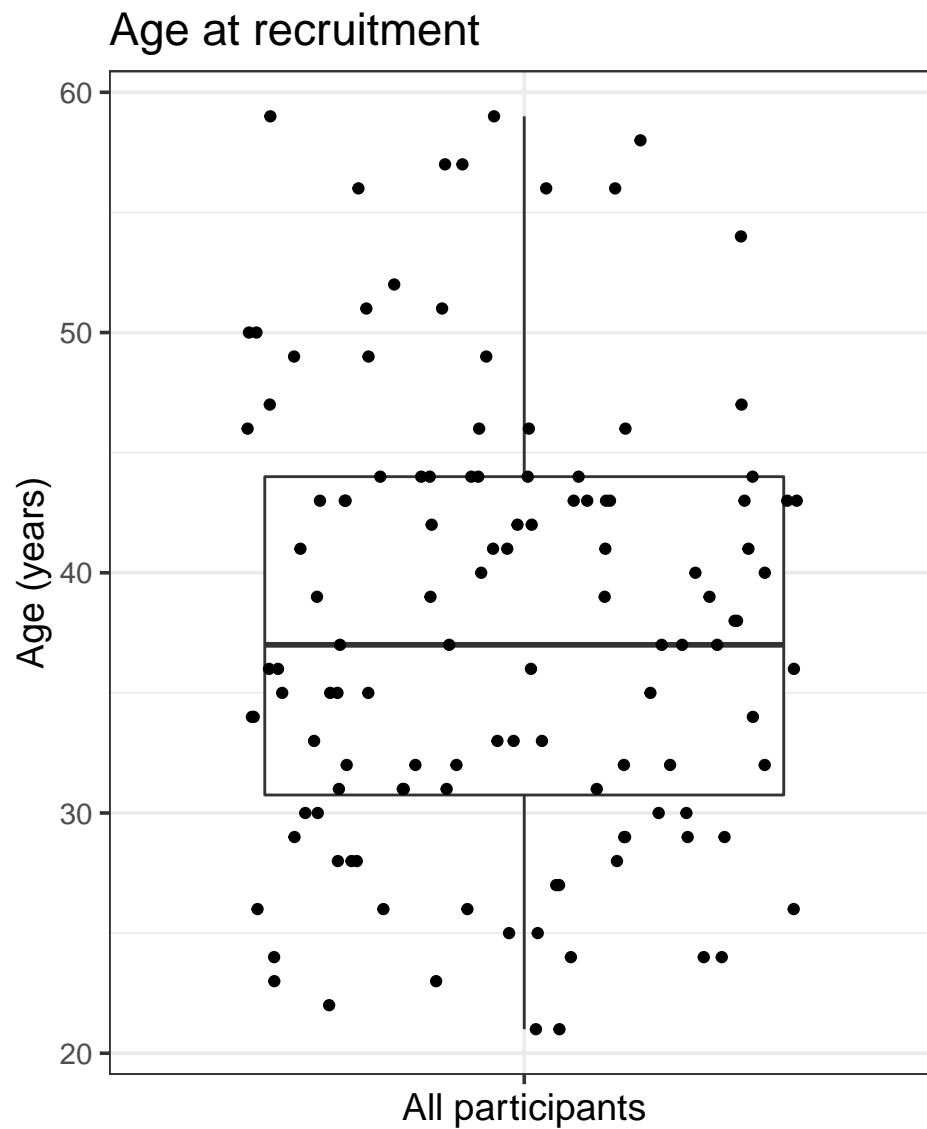
```r
# 95% bootstrap confidence interval of the mean age
## Method = BCa, Resamples = 1999
set.seed(1234)
groupwiseMean(age_years ~ 1,
              data = data,
              R = 1999,
              traditional = FALSE,
              boot = TRUE,
              bca = TRUE)[c(2:3, 5, 6, 7)]
```

```
##      n Mean Conf.level Bca.lower Bca.upper
## 1 120 37.8       0.95        36      39.5
```

```
# Plot
data %>%
    ggplot(data = .) +
    aes(y = age_years,
        x = 'All patients') +
    geom_boxplot() +
    geom_jitter(height = 0) +
    labs(title = 'Age at recruitment',
         y = 'Age (years)',
         x = 'All participants') +
    theme(axis.text.x = element_blank())
```
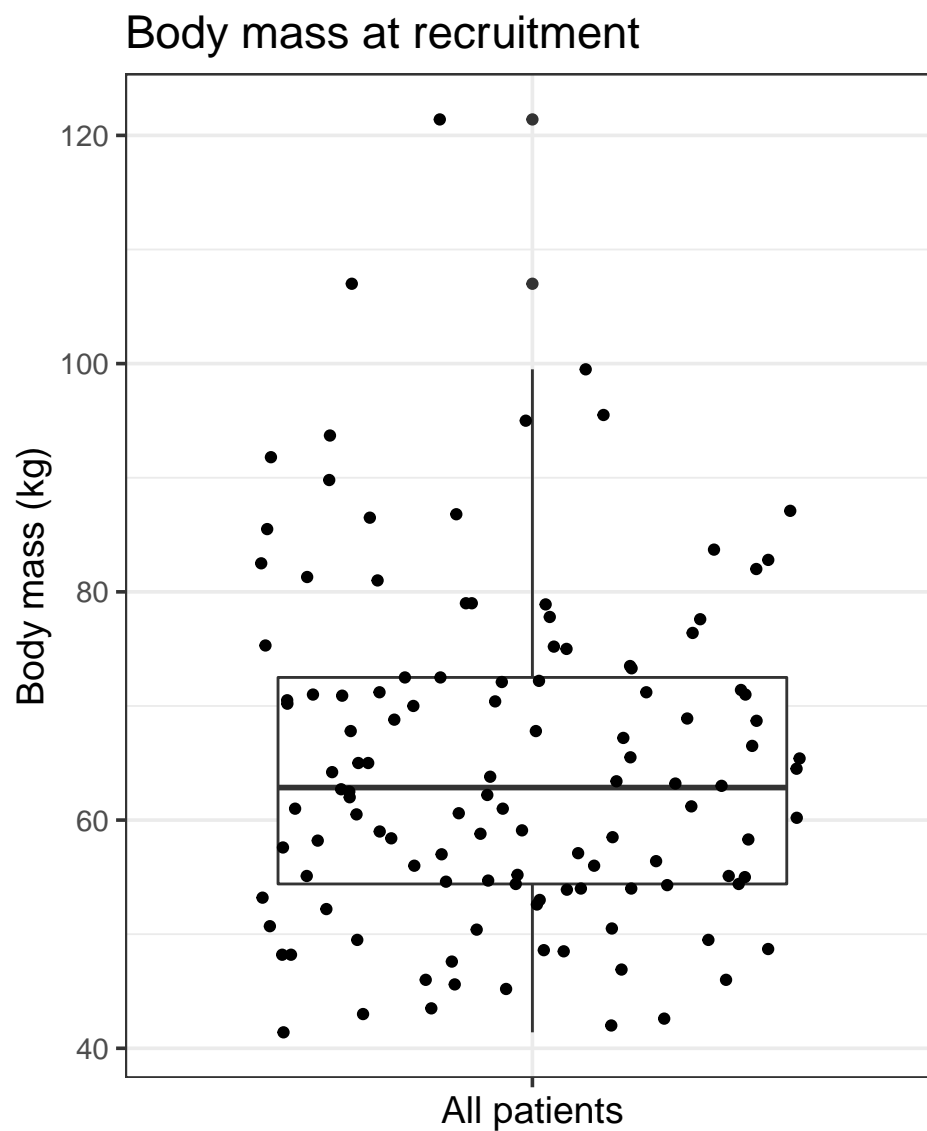

Age at recruitment

Body mass

```r
# Tabular summary
data %>%
    select(mass_kg) %>%
    skim()
```

```
## Skim summary statistics
##  n obs: 120
##  n variables: 1
##
## -- Variable type:numeric ----------------------------------------------------------
##  variable missing complete   n  mean    sd   p0 p25   p50  p75  p100
##   mass_kg       0      120 120 65.02 14.71 41.4 54.4 62.85 72.5 121.4
##      hist
##
```

```r
# 95% bootstrap confidence interval of thew mean age
## Method = BCa, Resamples = 1999
set.seed(1234)
groupwiseMean(mass_kg ~ 1,
              data = data,
              R = 1999,
              traditional = FALSE,
              boot = TRUE,
              bca = TRUE)[c(2:3, 5, 6, 7)]
```

```
##       n Mean Conf.level Bca.lower Bca.upper
## 1 120   65       0.95      62.8      67.7
```

```r
# Plot
data %>%
    ggplot(data = .) +
    aes(y = mass_kg,
        x = 'All patients') +
    geom_boxplot() +
    geom_jitter(height = 0) +
    labs(title = 'Body mass at recruitment',
         y = 'Body mass (kg)') +
    theme(axis.text.x = element_blank())
```

Body mass at recruitment

Height

Expect height to show sex difference, so analyse separately for males and females.

```r
# Tabular summary
data %>%
    group_by(sex) %>%
    select(height_m, sex) %>%
    skim()
```
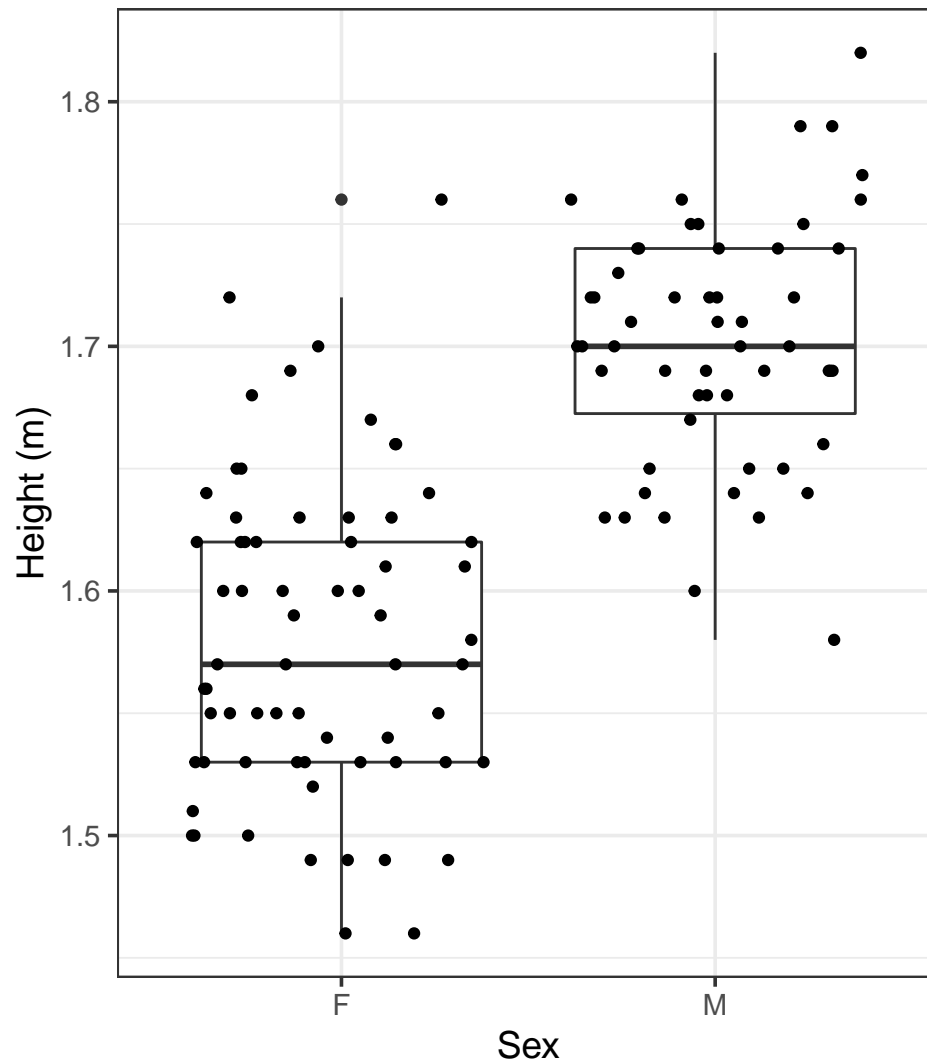
```
## Skim summary statistics
##  n obs: 120
##  n variables: 2
##  group variables: sex
##
## -- Variable type:numeric ----------------------------------------------------------------
##  sex variable missing complete  n mean     sd   p0  p25  p50  p75 p100
##    F height_m       0       66 66 1.58 0.065 1.46 1.53 1.57 1.62 1.76
```

```
##      M height_m        0        54 54 1.7   0.05   1.58 1.67 1.7   1.74 1.82
##        hist
##
##
```

```r
# 95% bootstrap confidence interval of the mean height
## Method = BCa, Resamples = 1999
set.seed(1234)
groupwiseMean(age_years ~ sex,
              data = data,
              R = 1999,
              traditional = FALSE,
              boot = TRUE,
              bca = TRUE)[c(1:3, 5, 6, 7)]
```

```
##   sex  n Mean Conf.level Bca.lower Bca.upper
## 1   F 66 36.0       0.95      33.8      38.5
## 2   M 54 39.9       0.95      37.7      42.2
```

```r
# Plots
data %>%
    ggplot(data = .) +
    aes(y = height_m,
        x = sex) +
    geom_boxplot() +
    geom_jitter(height = 0) +
    labs(title = 'Height at recruitment, grouped by sex',
         y = 'Height (m)',
         x = 'Sex')
```

Height at recruitment, grouped by sex

Sex

```
# Tabular summary
data %>%
    select(sex) %>%
    skim()
```

```
## Skim summary statistics
##  n obs: 120
##  n variables: 1
##
## -- Variable type:factor ------------------------------------------------------------------
##  variable missing complete   n n_unique          top_counts ordered
##       sex       0      120 120        2 F: 66, M: 54, NA: 0   FALSE
```

```
# 95% bootstrap confidence interval of the proportion of females
## Method = BCa, Resamples = 1999
set.seed(1234)
```

```r
boot.ci(boot(data = data,
             statistic = function(d, i){mean(d[i, 'sex'] == 'F')},
             R = 1999,
             stype = 'i'),
        type = 'bca') %>%
    tibble(n = nrow(filter(data, !is.na(sex))),
      Proportion = round(.$t0, 3),
      Conf.level = 0.95,
      Bca.lower = round(.$bca[[4]], 3),
      Bca.upper = round(.$bca[[5]], 3)) %>%
    .[1, -1] %>%
    as.data.frame()
```

```
##      n Proportion Conf.level Bca.lower Bca.upper
## 1 120       0.55       0.95     0.458     0.633
```

```r
# Plot
data %>%
    ggplot(data = .) +
    aes(x = 'All participants',
        fill = sex) +
    geom_bar(position = 'fill') +
    geom_hline(yintercept = 0.5,
               linetype = 2) +
    labs(title = 'Sex ratio at recruitment',
         subtitle = '(All participants)',
         y = 'Proportion') +
    scale_fill_grey(name = 'Sex') +
    theme(axis.text.x = element_blank())
```

Sex ratio at recruitment
(All participants)
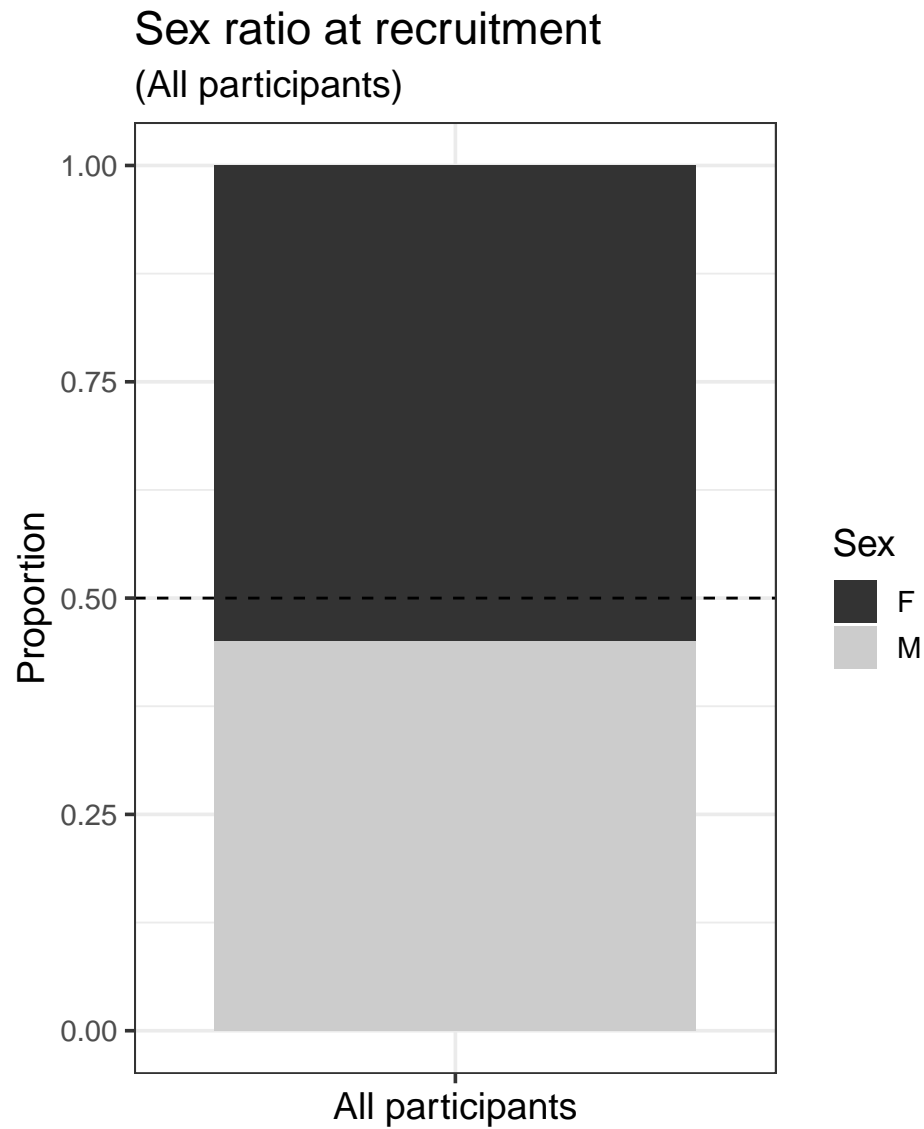
CD4 T-cell count

```
# Tabular summary
data %>%
    select(CD4_cell.ul) %>%
    skim()

## Skim summary statistics
##  n obs: 120
##  n variables: 1
##
## -- Variable type:numeric -------------------------------------------------------------
##     variable missing complete   n    mean      sd p0    p25 p50 p75 p100
##  CD4_cell.ul       1      119 120 265.87 224.75  1 119.5 221 345 1347
##      hist
##

# 95% bootstrap confidence interval of the median CD4 T-cell count
```

```
## Method = BCa, Resamples = 1999
set.seed(1234)
groupwiseMedian(CD4_cell.ul ~ 1,
                data = data[!is.na(data$CD4_cell.ul), ],
                R = 1999,
                boot = TRUE,
                bca = TRUE)[c(2:3, 5, 6, 7)]

##     n Median Conf.level Bca.lower Bca.upper
## 1 119    221       0.95       172       283

# Plot
data %>%
    filter(!is.na(CD4_cell.ul)) %>%
    ggplot(data = .) +
    aes(y = CD4_cell.ul,
        x = 'All patients') +
    geom_boxplot() +
    geom_jitter(height = 0) +
    labs(title = 'CD4 T-cell count at recruitment',
        y = expression(paste('CD4 T-cell count (cells.', mu, 'l^-1, ')'))) +
    theme(axis.text.x = element_blank())
```
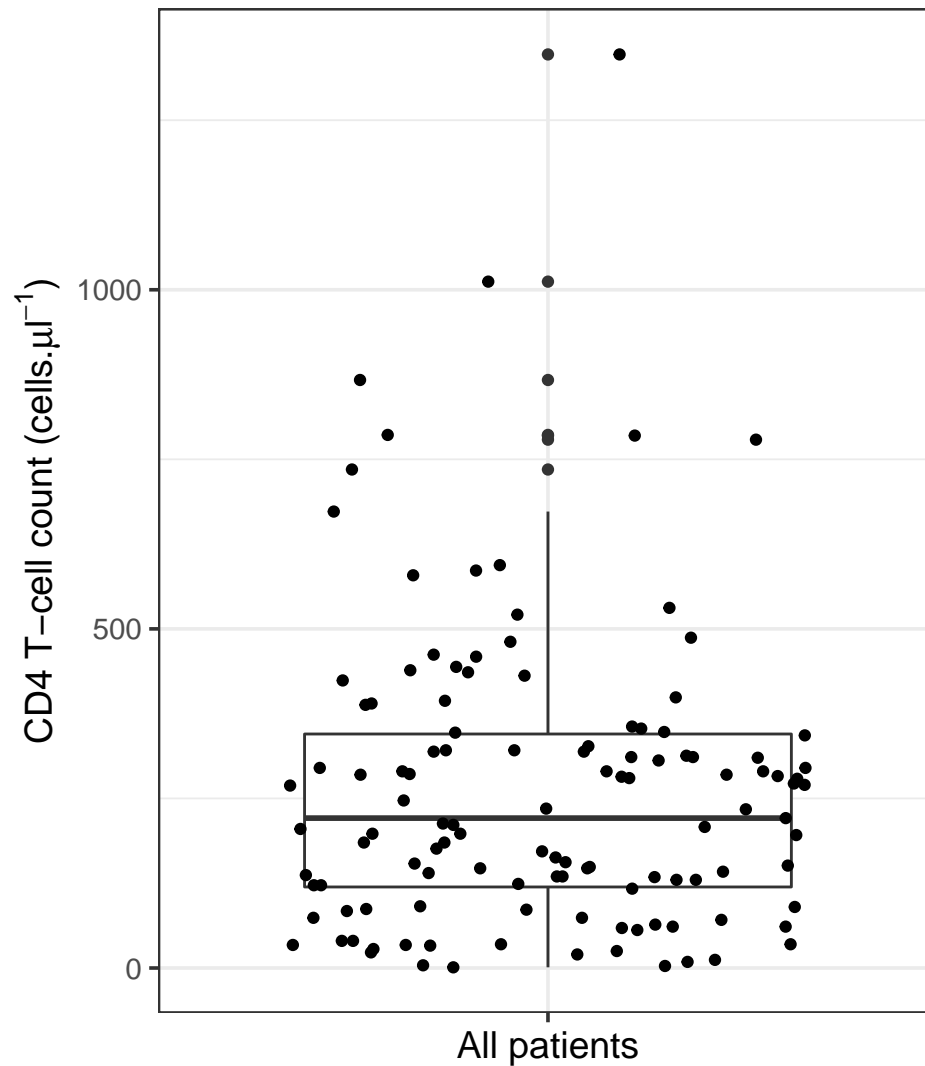
CD4 T−cell count at recruitment

Viral load

```r
# Tabular summary
data %>%
    select(viral_load_copies.ml) %>%
    skim()
```

```
## Skim summary statistics
##  n obs: 120
##  n variables: 1
##
## -- Variable type:numeric ------------------------------------------------------
##             variable missing complete   n mean   sd  p0 p25  p50  p75
##  viral_load_copies.ml      12      108 120 3.47 1.31 1.7 2.52 3.14 4.26
##  p100     hist
##  6.51
```

```r
# 95% bootstrap confidence interval of the median viral load
```

```
## Method = BCa, Resamples = 1999
set.seed(1234)
groupwiseMedian(viral_load_copies.ml ~ 1,
                data = data[!is.na(data$viral_load_copies.ml), ], # Remove <NA>
                R = 1999,
                boot = TRUE,
                bca = TRUE)[c(2:3, 5, 6, 7)]

##     n Median Conf.level Bca.lower Bca.upper
## 1 108   3.14       0.95      2.93       3.5

# Plot
data %>%
    filter(!is.na(viral_load_copies.ml)) %>%
    ggplot(data = .) +
    aes(y = viral_load_copies.ml,
        x = 'All patients') +
    geom_boxplot() +
    geom_jitter(height = 0) +
    labs(title = 'Viral load at recruitment',
         y = expression(paste('log' [10], ' viral load (copies.ml' ^-1, ')')))  +
    theme(axis.text.x = element_blank())
```

Viral load at recruitment
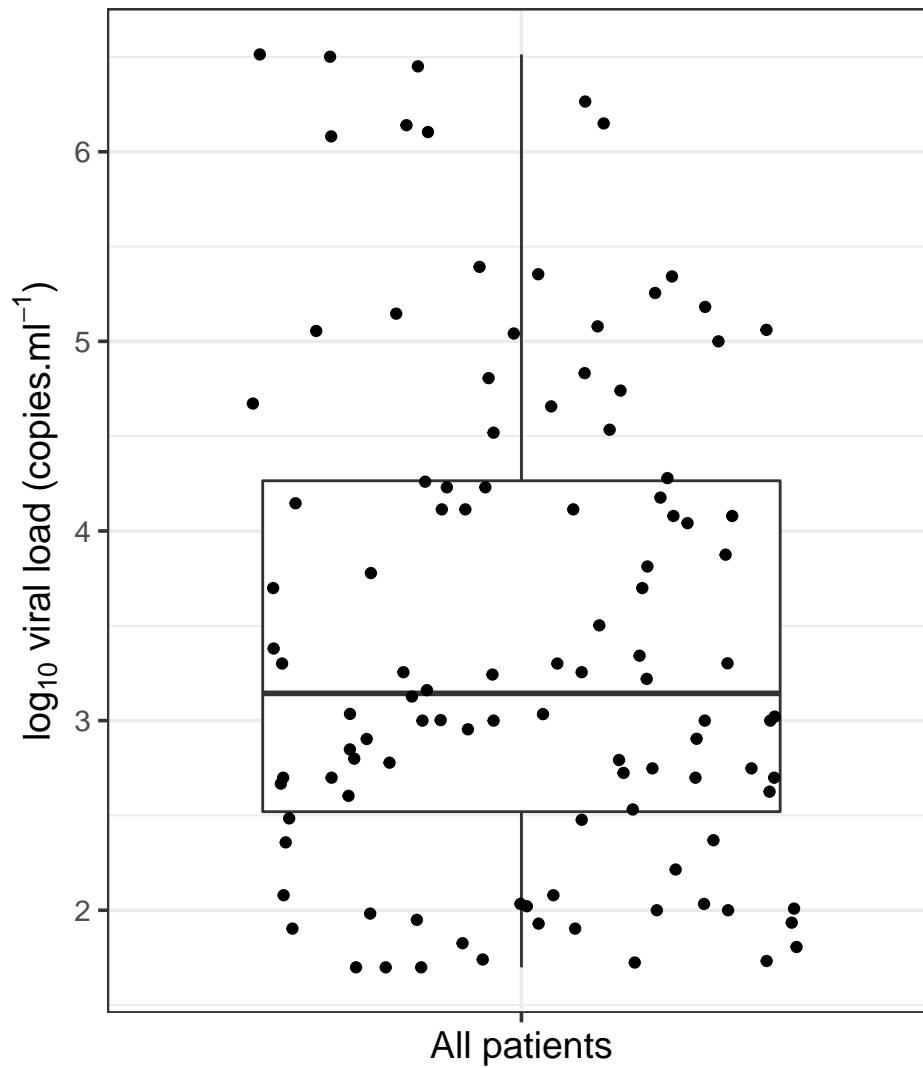
Alcohol

```
# Tabular summary
data %>%
    select(alcohol_units.week) %>%
    mutate(drinks_alcohol = case_when(
        alcohol_units.week >= 1 ~ 'Yes',
        alcohol_units.week == 0 ~ 'No'
    )) %>%
    mutate(drinks_alcohol = factor(drinks_alcohol)) %>%
    group_by(drinks_alcohol) %>%
    skim()
```

```
## Skim summary statistics
##  n obs: 120
##  n variables: 2
##  group variables: drinks_alcohol
```

```
## 
## -- Variable type:integer ------------------------------------------------------------------
##  drinks_alcohol           variable missing complete  n  mean    sd p0 p25
##             No alcohol_units.week       0       93 93 0     0     0   0
##            Yes alcohol_units.week       0       27 27 23.67 25.43  3   8
##  p50  p75 p100      hist
##    0  0      0
##   15 25.5   95
```

```r
# 95% bootstrap confidence interval of the median alcohol consumption
## Method = BCa, Resamples = 1999
set.seed(1234)
groupwiseMedian(alcohol_units.week ~ 1,
                data = data[data$alcohol_units.week > 0, ], # Remove none drinkers
                R = 1999,
                boot = TRUE,
                bca = TRUE)[c(2:3, 5, 6, 7)]
```

```
##    n Median Conf.level Bca.lower Bca.upper
## 1 27     15       0.95         6        18
```

```r
# Plot
data %>%
    filter(alcohol_units.week > 0) %>%
    ggplot(data = .) +
    aes(x = 'Participants who drink\n(n=16)',
        y = alcohol_units.week) +
    geom_boxplot() +
    geom_jitter(height = 0) +
    labs(title = 'Alcohol consumption at recruitment',
         y = expression(paste('Alcohol consumption (units.week' ^-1, ')'))) +
    theme(axis.text.x = element_blank())
```

Alcohol consumption at recruitment

Participants who drink
(n=16)
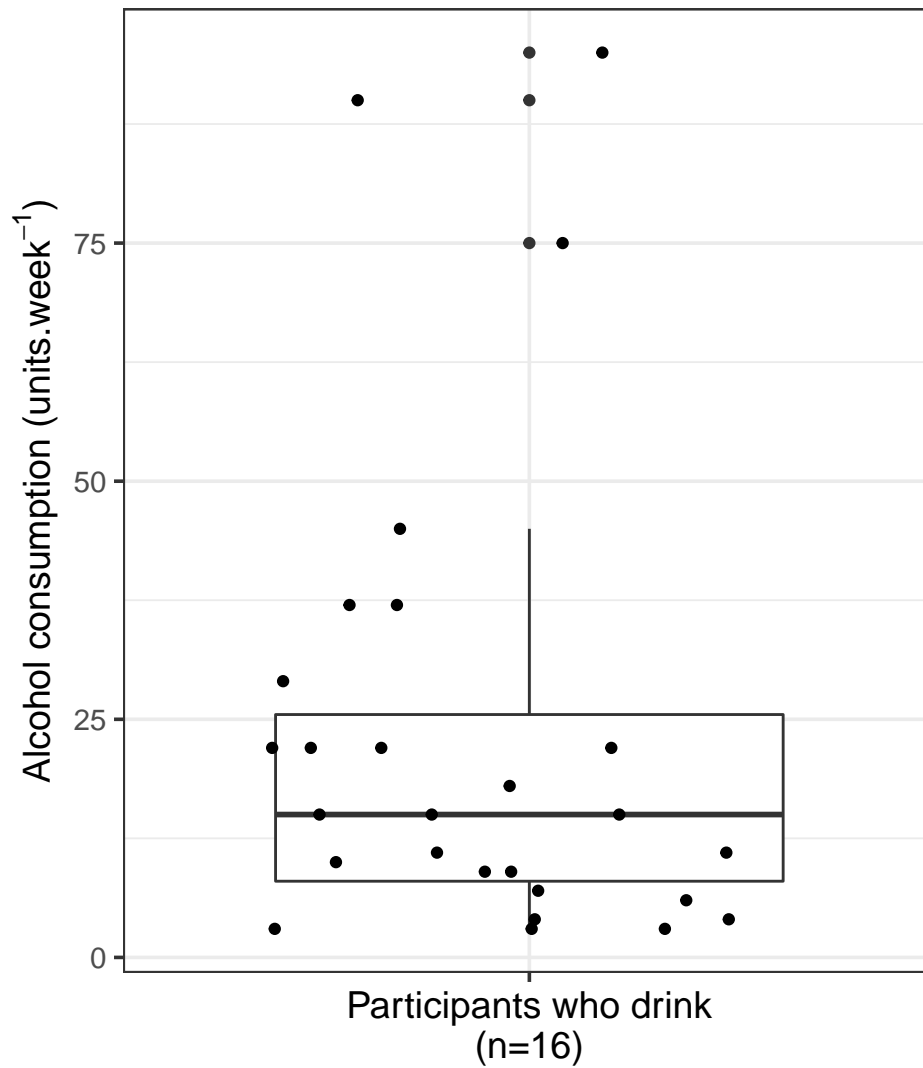
TB

Note: Treatment policy was to start some patients, irrespective of TB diagnosis, on TB treatment. Therefore current TB infection and treatment for TB analysed separately.

Currently infected with TB

```r
# Tabular summary
data %>%
    select(TB_current) %>%
    skim()
```

```
## Skim summary statistics
##  n obs: 120
##  n variables: 1
##
## -- Variable type:factor --------------------------------------------------------
```

```
##     variable missing complete   n n_unique             top_counts ordered
##  TB_current       0      120 120        2 no: 100, yes: 20, NA: 0   FALSE
```

```r
# 95% bootstrap confidence interval of the proportion with TB
## Method = BCa, Resamples = 1999
set.seed(1234)
boot.ci(boot(data = data,
             statistic = function(d, i){mean(d[i, 'TB_current'] == 'yes')},
             R = 1999,
             stype = 'i'),
        type = 'bca') %>%
    tibble(n = nrow(filter(data, !is.na(TB_current))),
           Proportion = round(.$t0, 3),
           Conf.level = 0.95,
           Bca.lower = round(.$bca[[4]], 3),
           Bca.upper = round(.$bca[[5]], 3)) %>%
    .[1, -1] %>%
    as.data.frame()
```

```
##     n Proportion Conf.level Bca.lower Bca.upper
## 1 120      0.167       0.95     0.108     0.233
```

```r
# Plot
data %>%
    mutate(TB_current = str_to_title(TB_current)) %>%
    ggplot(data = .) +
    aes(x = 'All participants',
        fill = TB_current) +
    geom_bar(position = 'fill') +
    geom_hline(yintercept = 0.5,
               linetype = 2) +
    labs(title = 'Current TB at recruitment',
         y = 'Proportion') +
    scale_fill_grey(name = 'Infected') +
    theme(axis.text.x = element_blank())
```

Current TB at recruitment
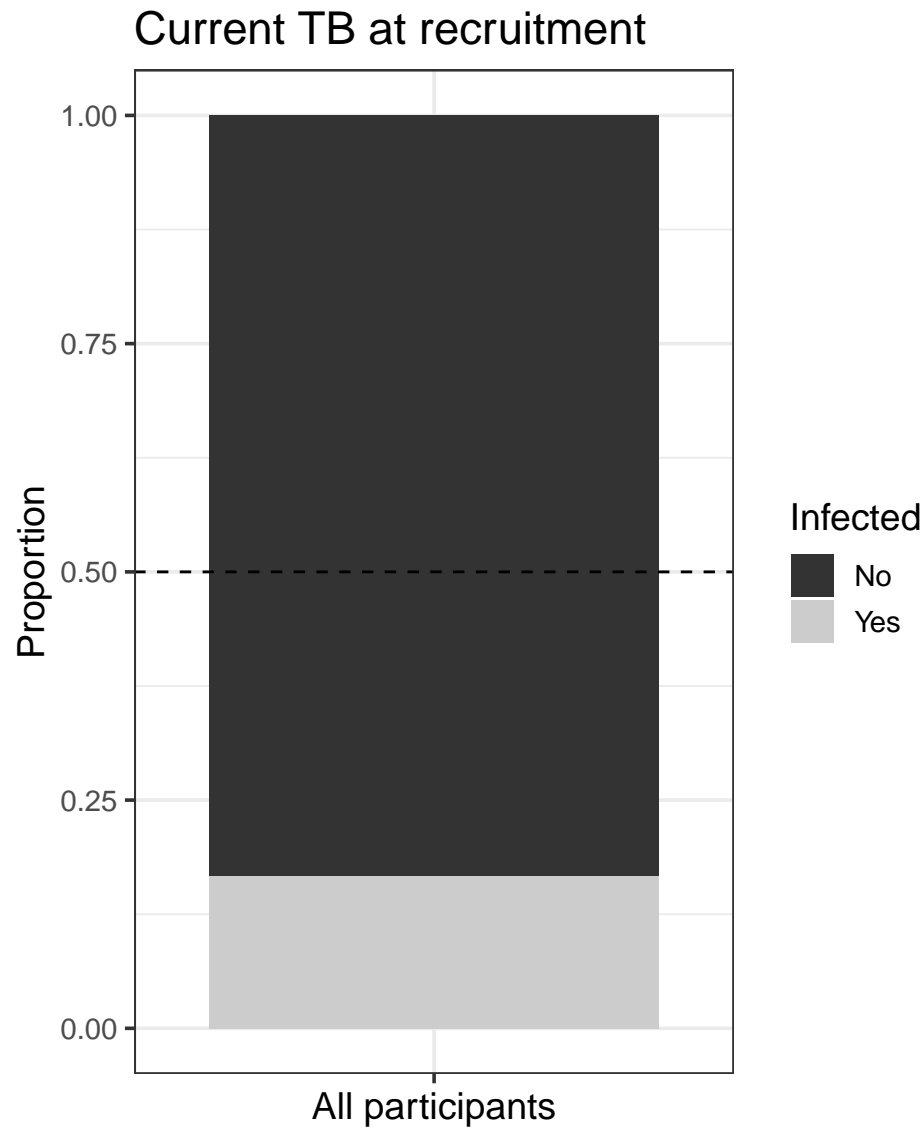
Currently receiving TB treatment?

Treatment consisted of rifafour and pyridoxine (prophylaxis). Therefore only need to analyse rifafour data. Data coded as *'No'* (not being treated), *'Yes'* (being treated for active TB), and *'Prophylaxis'* (being treated prophylactically for TB).

```
# Double-check matching between rifafour and pyridoxine columns
unique(data$rifafour_treatment == data$pyridoxine_prophylaxis)
```

```
## [1] TRUE
```

```
# Tabular summary
data %>%
    group_by(pyridoxine_prophylaxis) %>%
    select(rifafour_treatment, pyridoxine_prophylaxis) %>%
    skim()
```

```
## Skim summary statistics
##  n obs: 120
```

```
##  n variables: 2
##  group variables: pyridoxine_prophylaxis
##
## -- Variable type:factor ----------------------------------------------------------------
##  pyridoxine_prophylaxis           variable missing complete  n n_unique
##                      no rifafour_treatment       0       87 87        1
##                     yes rifafour_treatment       0       19 19        1
##             prophylaxis rifafour_treatment       0       14 14        1
##                   top_counts ordered
##  no: 87, yes: 0, pro: 0, NA: 0   FALSE
##  yes: 19, no: 0, pro: 0, NA: 0   FALSE
##  pro: 14, no: 0, yes: 0, NA: 0   FALSE
```

```r
# Proportion on prophylaxis treatment
## Too low to analyse separately
round(mean(data$rifafour_treatment == 'prophylaxis'), 3)
```

```
## [1] 0.117
```

```r
## ...so collapse 'yes' and 'prophylaxis'
data_tb <- data %>%
    mutate(rifafour_treatment = fct_collapse(rifafour_treatment,
                                             yes = c('yes', 'prophylaxis')))
```

```r
# 95% bootstrap confidence interval of the proportion on TB treatment
## Method = BCa, Resamples = 1999
set.seed(1234)
boot.ci(boot(data = data_tb,
             statistic = function(d, i){mean(d[i, 'rifafour_treatment'] == 'yes')},
             R = 1999,
             stype = 'i'),
        type = 'bca') %>%
    tibble(n = nrow(filter(data_tb, !is.na(rifafour_treatment))),
           Proportion = round(.$t0, 3),
           Conf.level = 0.95,
           Bca.lower = round(.$bca[[4]], 3),
           Bca.upper = round(.$bca[[5]], 3)) %>%
    .[1, -1] %>%
    as.data.frame()
```

```
##     n Proportion Conf.level Bca.lower Bca.upper
## 1 120      0.275       0.95     0.195     0.358
```

```r
# Plot
data %>%
    mutate(rifafour_treatment = str_to_title(rifafour_treatment),
           rifafour_treatment = factor(rifafour_treatment,
                                       levels = c('No', 'Yes',
                                                  'Prophylaxis'),
                                       ordered = TRUE)) %>%
    ggplot(data = .) +
    aes(x = 'All participants',
        fill = rifafour_treatment) +
    geom_bar(position = 'fill') +
    geom_hline(yintercept = 0.5,
               linetype = 2) +
```

```
labs(title = 'Being treated for TB at recruitment',
     y = 'Proportion') +
scale_fill_grey(name = 'Treatment') +
theme(axis.text.x = element_blank())
```

Being treated for TB at recruitment
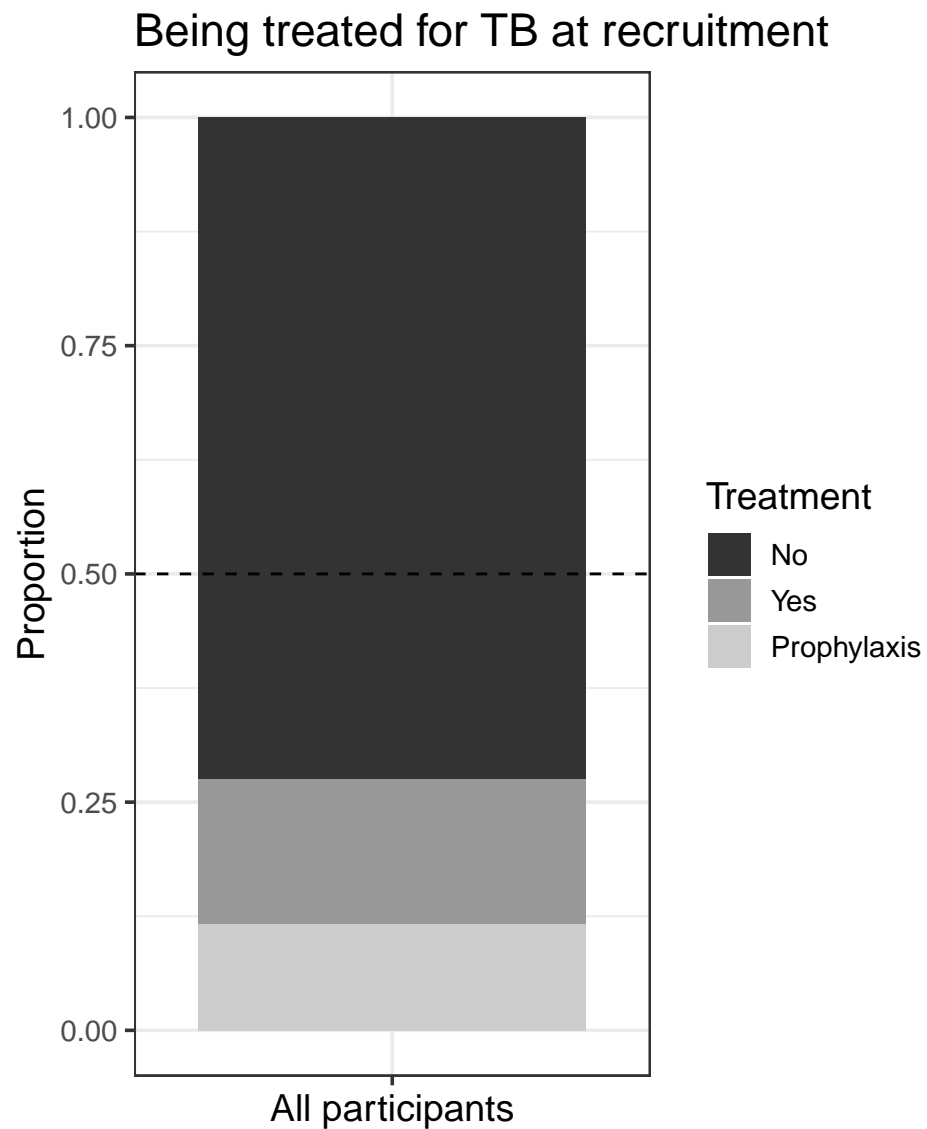
Diabetes

Classified as diabetic based on `data$hba1c_percent` > 7%. No participants were diabetic.

```
# Tabular summary
data %>%
    select(diabetic_hba1c) %>%
    skim()
```

```
## Skim summary statistics
##  n obs: 120
##  n variables: 1
##
```

```
## -- Variable type:factor ----------------------------------------------------------------------
##         variable missing complete   n n_unique             top_counts
##   diabetic_hba1c       9      111 120        1 no: 111, NA: 9, yes: 0
##  ordered
##    FALSE
```

Vitamin B12 deficiency

Classed as B12 deficient based on `data$vitaminB12_pmol.l` < 141 pmol/l. Only one participant had a deficiency.

```
# Tabular summary
data %>%
    select(vitaminB12_deficiency) %>%
    skim()

## Skim summary statistics
##  n obs: 120
##  n variables: 1
##
## -- Variable type:factor ----------------------------------------------------------------------
##               variable missing complete   n n_unique
##   vitaminB12_deficiency      19      101 120        2
##               top_counts ordered
##   no: 100, NA: 19, yes: 1   FALSE
```

Session information

```
sessionInfo()

## R version 3.5.2 (2018-12-20)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Mojave 10.14.3
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] skimr_1.0.5       boot_1.3-20       rcompanion_2.1.1 forcats_0.4.0
##  [5] stringr_1.4.0     dplyr_0.8.0.1     purrr_0.3.1       readr_1.3.1
##  [9] tidyr_0.8.3       tibble_2.0.1      ggplot2_3.1.0     tidyverse_1.2.1
## [13] magrittr_1.5
##
## loaded via a namespace (and not attached):
```

```
##  [1] Rcpp_1.0.0          lubridate_1.7.4    mvtnorm_1.0-9
##  [4] lattice_0.20-38     multcompView_0.1-7 zoo_1.8-4
##  [7] utf8_1.1.4          lmtest_0.9-36      assertthat_0.2.0
## [10] digest_0.6.18       R6_2.4.0           cellranger_1.1.0
## [13] plyr_1.8.4          backports_1.1.3    EMT_1.1
## [16] stats4_3.5.2        evaluate_0.13      httr_1.4.0
## [19] pillar_1.3.1        rlang_0.3.1        lazyeval_0.2.1
## [22] multcomp_1.4-8      readxl_1.3.0       rstudioapi_0.9.0
## [25] Matrix_1.2-15       rmarkdown_1.11     labeling_0.3
## [28] splines_3.5.2       foreign_0.8-71     munsell_0.5.0
## [31] broom_0.5.1         compiler_3.5.2     modelr_0.1.4
## [34] xfun_0.5            pkgconfig_2.0.2    manipulate_1.0.1
## [37] DescTools_0.99.27   htmltools_0.3.6    tidyselect_0.2.5
## [40] expm_0.999-3        coin_1.2-2         codetools_0.2-16
## [43] fansi_0.4.0         crayon_1.3.4       withr_2.1.2.9000
## [46] MASS_7.3-51.1       grid_3.5.2         nlme_3.1-137
## [49] jsonlite_1.6        gtable_0.2.0       scales_1.0.0
## [52] cli_1.0.1           stringi_1.3.1      xml2_1.2.0
## [55] generics_0.0.2      nortest_1.0-4      sandwich_2.5-0
## [58] TH.data_1.0-10      tools_3.5.2        glue_1.3.0
## [61] hms_0.4.2           survival_2.43-3    yaml_2.2.0
## [64] colorspace_1.4-0    rvest_0.3.2        knitr_1.21
## [67] haven_2.1.0         modeltools_0.2-22
```