Biochemistry

Nathan Constantine-Cooke

2025-11-07

Table of contents

Introduction

```
set.seed(123)

source("Baseline/utils.R")

###############
## Packages ##
###############

library(plyr) # Used for mapping values
suppressPackageStartupMessages(library(tidyverse)) # ggplot2, dplyr, and magrittr
library(readxl) # Read in Excel files
library(lubridate) # Handle dates
library(datefixR) # Standardise dates
library(patchwork) # Arrange ggplots
```

```r
# Generate tables
suppressPackageStartupMessages(library(table1))
library(knitr)
library(pander)

# Generate flowchart of cohort derivation
library(DiagrammeR)
library(DiagrammeRsvg)

# paths to PREdiCCt data
if (file.exists("/docker")) { # If running in docker
  data.path <- "data/final/20221004/"
  redcap.path <- "data/final/20231030/"
  prefix <- "data/end-of-follow-up/"
  outdir <- "data/processed/"
} else { # Run on OS directly
  data.path <- "/Volumes/igmm/cvallejo-predicct/predicct/final/20221004/"
  redcap.path <- "/Volumes/igmm/cvallejo-predicct/predicct/final/20231030/"
  prefix <- "/Volumes/igmm/cvallejo-predicct/predicct/end-of-follow-up/"
  outdir <- "/Volumes/igmm/cvallejo-predicct/predicct/processed/"
}

demo <- readRDS(paste0(outdir, "demo-IBD.RDS"))

labs <- as.data.frame(read_xlsx(paste0(
  data.path,
  "Baseline2022/bloodtestresults.xlsx"
)))

labs <- labs %>%
  distinct(ParticipantNo, .keep_all = TRUE)


labs <- labs[, c(
  "ParticipantNo",
  "CReactiveProtein",
  "Haemoglobin",
  "HaemoglobinUnit",
  "WCC",
  "WCCUnit",
  "Platelets",
  "PlateletsUnit",
```

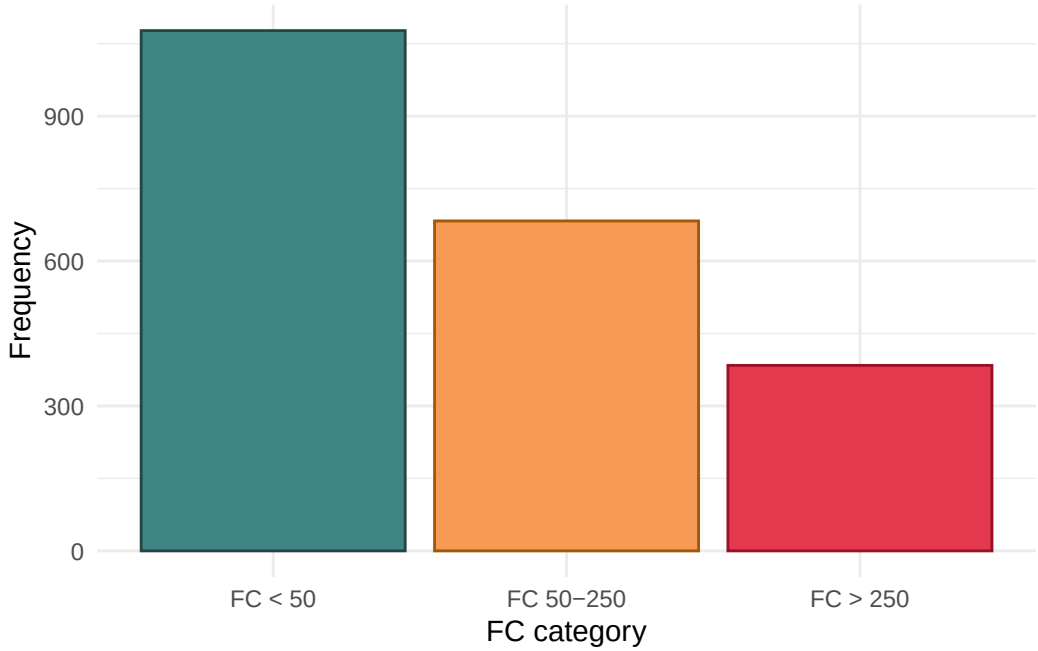
```
  "Albumin",
  "AlbuminUnit"
)]
demo <- merge(demo, labs, by = "ParticipantNo", all.x = TRUE, all.y = FALSE)
```

Participants' IBD care teams reported biochemistry results via REDCap. Whilst data from many types of blood tests were collected, only the most relevant tests are presented here. Where possible, reference levels have been denoted on plots via vertical lines.

Faecal calprotectin

As a reminder, only subjects with a faecal calprotectin are included in this analysis, and therefore NA values are not reported for faecal calprotectin. Faecal calprotectin has been grouped into FC<50, 50 FC 250, and FC > 250.

```
demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = cat, color = cat, fill = cat)) +
  geom_bar() +
  labs(x = "FC category", y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("#408585", "#F59954", "#E4384E")) +
  scale_color_manual(values = c("#244644", "#A35805", "#9A1027"))
```



C-reactive protein

```
perc <- (1 - demo %>%
  drop_na(cat) %>%
  select(CReactiveProtein) %>%
  is.na() %>%
  sum() /
  demo %>%
    drop_na(cat) %>%
    nrow()) * 100
```

Alongside FC, CRP is the most requested laboratory test for monitoring IBD disease activity. Unlike FC which acts as a proxy for gastrointestinal inflammation, CRP provides an indication of inflammation across the body. The reference level for CRP is considered to be < 5 mg/L.

CRP was generally well-completed with 76.4% of subjects having an associated CRP. As can be seen in Figure 1, there were some extreme, albeit physiologically plausible, values reported.

```
demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = CReactiveProtein)) +
```

```
  geom_histogram(bins = 100, fill = "#EB9486", color = "#BF7163") +
  theme_minimal() +
  xlab("CRP (mg/L)") +
  ylab("Frequency")
```

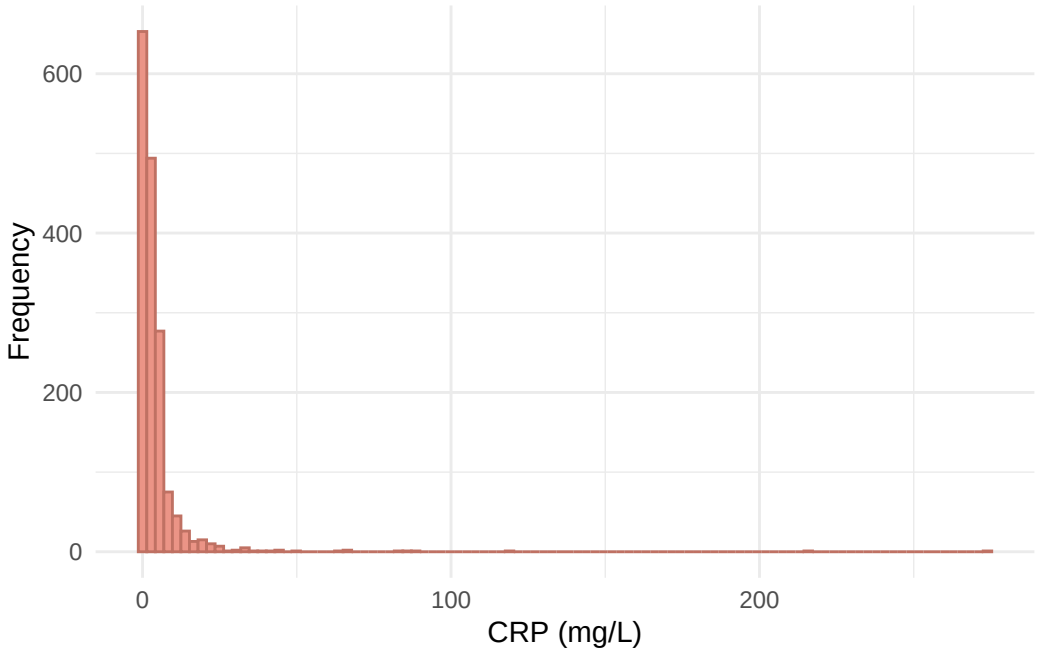


Figure 1: Distribution of CRP at recruitment for the FC cohort.

As should be expected, CRP was found to be associated with FC.

```
pander(summary(aov(CReactiveProtein ~ cat, data = demo)))
```

Table 1: ANOVA between CRP and FC groups.

Table 1: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat	2	1949	974.4	7.98	0.0003556
Residuals	1635	199628	122.1	NA	NA

```r
p <- demo %>%
  drop_na(cat) %>%
  ggplot(aes(x=cat, y = CReactiveProtein)) +
  geom_boxplot(staplewidth  = 1,
               fill = "#C0D8E0",
               color = "#42797B",
               outlier.shape = NA) +
  ylim(0, 12) +
  theme_minimal() +
  xlab("Faecal calprotectin category (\U03BCg/g)") +
  ylab("C-reactivive protein (mg/L)")
cairo_pdf("plots/CRP-FC-boxplot.pdf", width = 8, height = 6)
p
```

```
Warning: Removed 601 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

```r
invisible(dev.off())

p
```

```
Warning: Removed 601 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

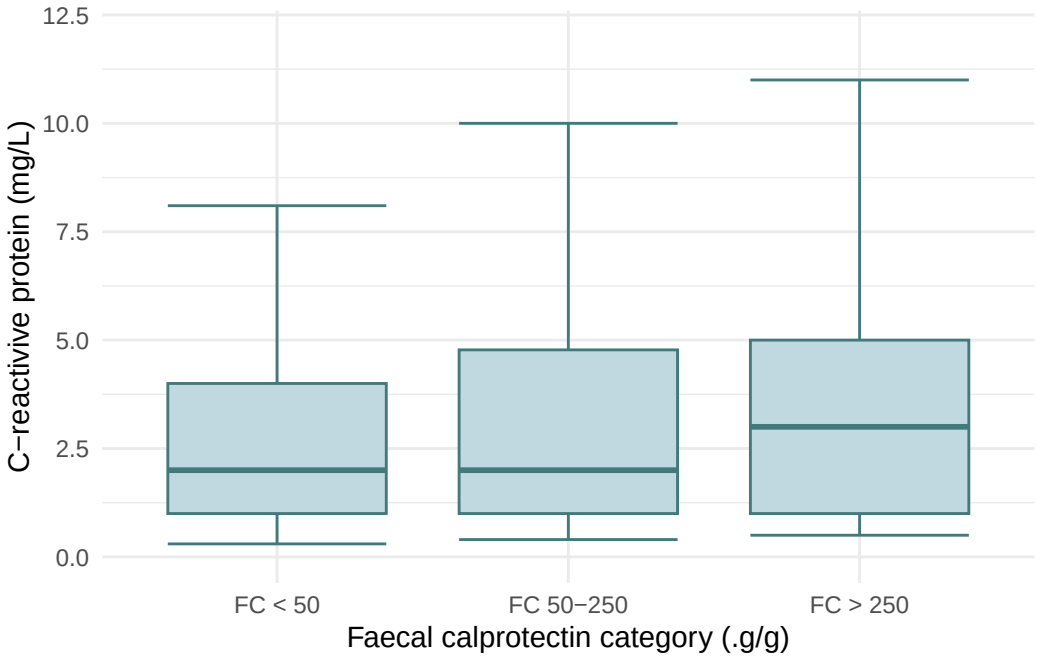


Figure 2: Boxplot of CRP by FC category.

Haemoglobin

```
perc <- (1 - demo %>%
  drop_na(cat) %>%
  select(Haemoglobin) %>%
  is.na() %>%
  sum() /
  demo %>%
    drop_na(cat) %>%
    nrow()) * 100
```

Haemoglobin is a protein which facilitates the transport of oxygen in red blood cells. Haemoglobin has been reported for 83.07% of subjects in the FC cohort. Haemoglobin differs between sex and accordingly has sex-specific reference levels (130-170 g/L for men and 120-150g/L for women).

Whilst a normal distribution for haemoglobin is observed overall (Figure 3), there appears to be some observations substantially lower than one would expect for this distribution. Investigating these observations suggests that these observations are likely to be in units of g/dL instead of the common g/L units (Table 2).

```r
demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = Haemoglobin)) +
  geom_histogram(bins = 100, fill = "#76E5FC", color = "#2EB6CD") +
  theme_minimal() +
  xlab("Haemoglobin") +
  ylab("Frequency")
```

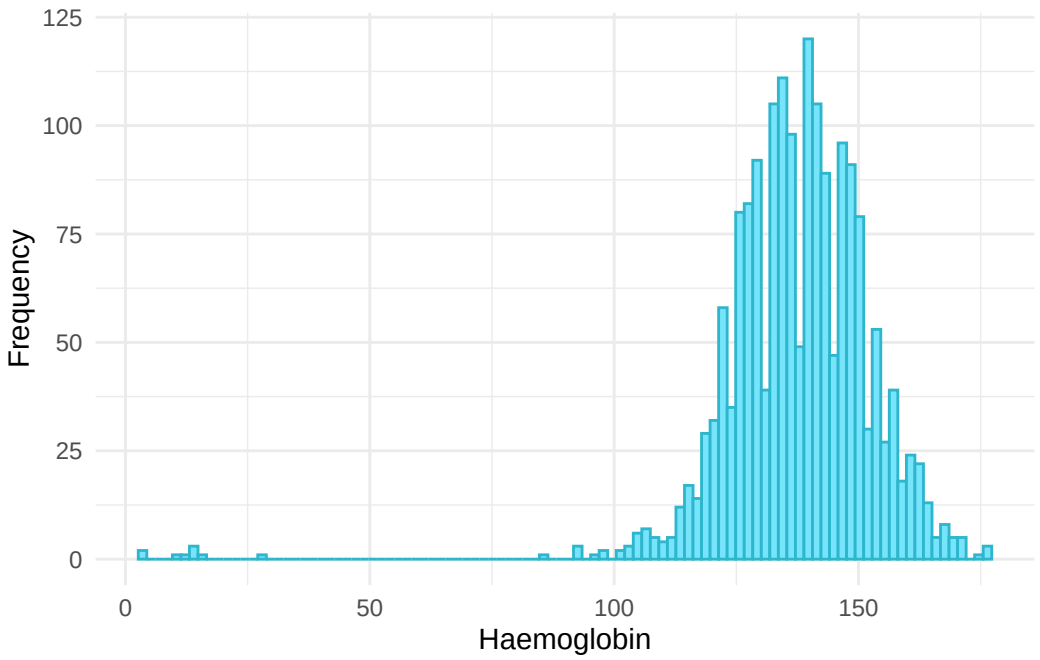


Figure 3: Distribution of haemoglobin before processing.

```r
knitr::kable(table(subset(demo, Haemoglobin < 50)$HaemoglobinUnit),
  col.names = c("Units", "Frequency")
)
```

Table 2: Reported unit measurements for Haemoglobin reported less than 50.

Units	Frequency
g/dL	6
g/l	1
g/L	5
mmol/mol	1

Table 2: Reported unit measurements for Haemoglobin reported less than 50.

Units	Frequency

Having assumed any Haemoglobin < 50 has been reported in g/dL, any Haemoglobin observations <50 have been multiplied by 10 to be in g/L units. This results in the distribution seen in Figure 4.

```
demo <- demo %>%
  mutate(Haemoglobin = if_else(Haemoglobin < 50,
    Haemoglobin * 10,
    Haemoglobin
  )) %>%
  select(-HaemoglobinUnit)

p1 <- demo %>%
  drop_na(cat) %>%
  filter(Sex == "Female") %>%
  ggplot(aes(x = Haemoglobin)) +
  geom_histogram(bins = 100, fill = "#8789C0", color = "#6A6B9C") +
  theme_minimal() +
  xlab("Haemoglobin (g/L)") +
  ylab("Frequency") +
  xlim(0, 190) +
  geom_vline(xintercept = c(120, 150), color = "#93032E") +
  ggtitle("Female haemoglobin")



p2 <- demo %>%
  drop_na(cat) %>%
  filter(Sex == "Male") %>%
  ggplot(aes(x = Haemoglobin, fill = Sex, color = Sex)) +
  geom_histogram(bins = 100, fill = "#225560", color = "#044651") +
  theme_minimal() +
  xlab("Haemoglobin (g/L)") +
  ylab("Frequency") +
  xlim(0, 190) +
  geom_vline(xintercept = c(130, 170), color = "#93032E") +
  ggtitle("Male haemoglobin")

(p1 / p2)
```

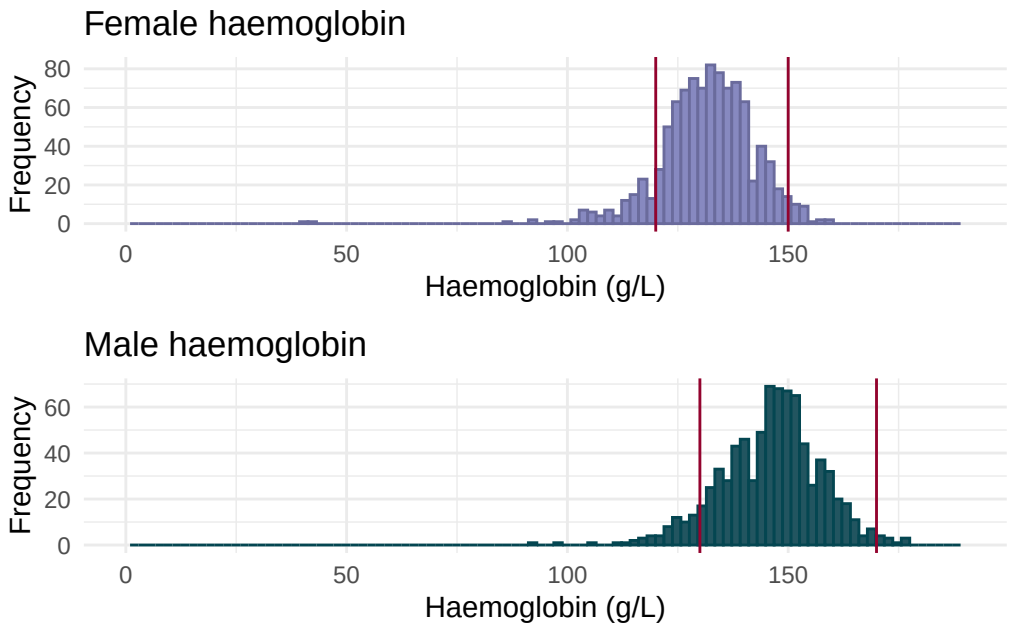


Figure 4: Distribution of haemoglobin after processing.

Haemoglobin was not found to be significantly associated with FC at a 5% significance level.

```
pander(summary(aov(Haemoglobin ~ cat, data = demo)))
```

Table 3: ANOVA between haemoglobin and FC groups.

Table 3: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat	2	940.6	470.3	2.539	0.07921
Residuals	1778	329292	185.2	NA	NA

White cell count

```r
perc <- (1 - demo %>%
  drop_na(cat) %>%
  select(WCC) %>%
  is.na() %>%
  sum() /
  demo %>%
    drop_na(cat) %>%
    nrow()) * 100
```

White cell count (WCC) is a measurement of all white blood cells (neutrophils, lymphocytes, monocytes, basophils and eosinophils) which help to fight infections. 83.02% of subjects have a WCC recorded.

Whilst some extreme WCC values are observed, most observations fall within the reference range (4-11 (×10^9/L)) and a conventional bell curve is seen overall.

```r
demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = WCC)) +
  geom_histogram(bins = 100, fill = "#7FD8BE", color = "#2D977E") +
  theme_minimal() +
  xlab("White cell count (×10^9/L") +
  ylab("Frequency") +
  geom_vline(xintercept = c(4, 11), color = "#FF686B", alpha = 0.8)
```

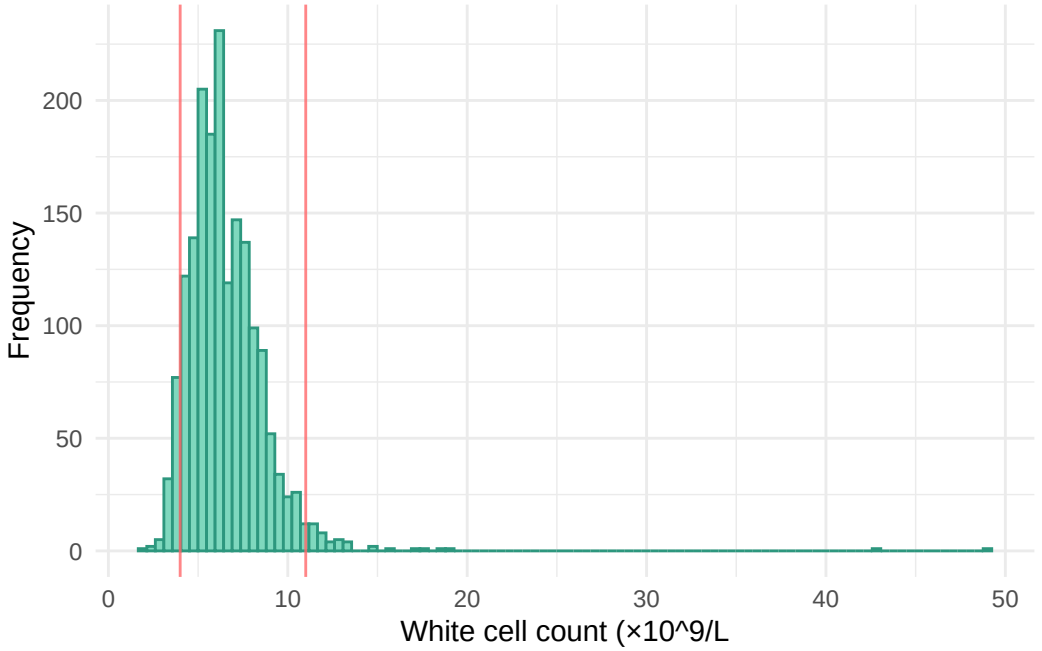


Figure 5: Distrbution of WCC for the FC cohort. Vertical lines indicate the reference range.

WCC is reported to be associated with FC.

```
pander(summary(aov(WCC ~ cat, data = demo)))
```

Table 4: ANOVA between WCC and FC groups.

Table 4: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat	2	95.27	47.63	8.615	0.0001891
Residuals	1777	9826	5.529	NA	NA

Platelets

```
perc <- (1 - demo %>%
  drop_na(cat) %>%
  select(Platelets) %>%
  is.na() %>%
```

```
  sum() /
  demo %>%
    drop_na(cat) %>%
    nrow()) * 100
```

The majority of subjects (83.02%) have a platelets test result recorded.

Whilst some extreme values are observed, most fall within the reference range.

```
demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = Platelets)) +
  geom_histogram(bins = 100, fill = "#84E6F8", color = "#018897") +
  theme_minimal() +
  xlab("Platelets (×10^9/L") +
  ylab("Frequency") +
  geom_vline(xintercept = c(150, 450), color = "#7b1907", alpha = 0.8)
```

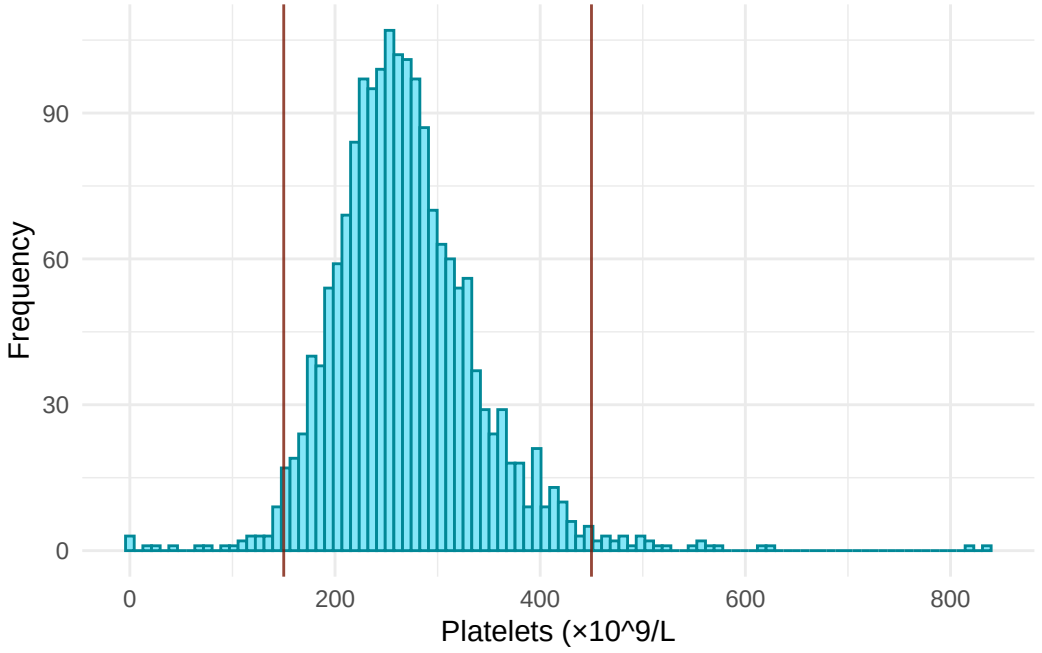


Figure 6: Distrbution of Platelets for the FC cohort. Vertical lines indicate the reference range.

Platelets is reported to be significantly associated with FC.

```
pander(summary(aov(Platelets ~ cat, data = demo)))
```

Table 5: ANOVA between platelets and FC groups.

Table 5: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat	2	96868	48434	9.635	6.891e-05
Residuals	1777	8933162	5027	NA	NA

Albumin

```
perc <- (1 - demo %>%
  drop_na(cat) %>%
  select(Albumin) %>%
  is.na() %>%
  sum() /
  demo %>%
    drop_na(cat) %>%
    nrow()) * 100
```

Albumin is globular protein made in the liver which carries hormones, vitamins, and enzymes. Albumin has a reference range of 35-50 (g/L). 79.1% of subjects in the FC cohort have a recorded Albumin result.

```
demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = Albumin)) +
  geom_histogram(bins = 100, fill = "#21D19F", color = "#26A37D") +
  theme_minimal() +
  xlab("Albumin") +
  ylab("Frequency") +
  geom_vline(xintercept = c(35, 50), color = "#FF686B")
```

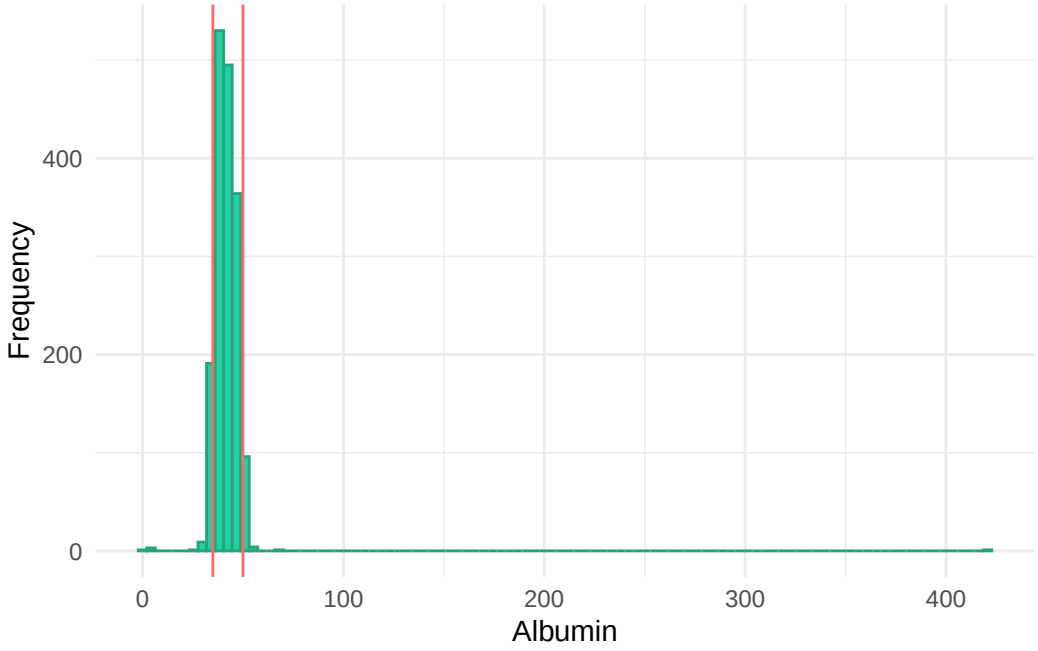


Figure 7: Distribution of Albumin before processing.

As can be seen in Figure 7, a very extreme value has been reported. This has been assumed to have been off by a factor of 10. Adjusting this observation results in the distribution seen in Figure 8.

```
demo <- demo %>%
  mutate(Albumin = if_else(Albumin > 100, Albumin / 10, Albumin)) %>%
  select(-AlbuminUnit)
demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = Albumin)) +
  geom_histogram(bins = 40, fill = "#21D19F", color = "#26A37D") +
  theme_minimal() +
  xlab("Albumin (g/L)") +
  ylab("Frequency") +
  geom_vline(xintercept = c(35, 50), color = "#FF686B")
```

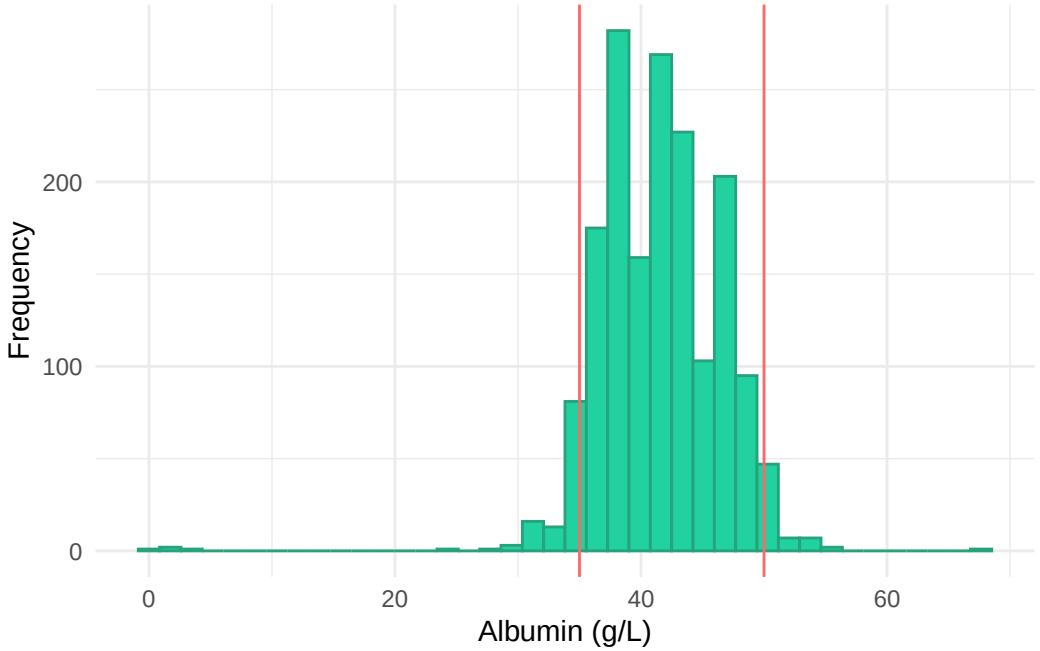


Figure 8: Distribution of Albumin after processing.

There is strong evidence for an association between Albumin and FC.

```
pander(summary(aov(Albumin ~ cat, data = demo)))
```

Table 6: ANOVA between Albumin and FC groups.

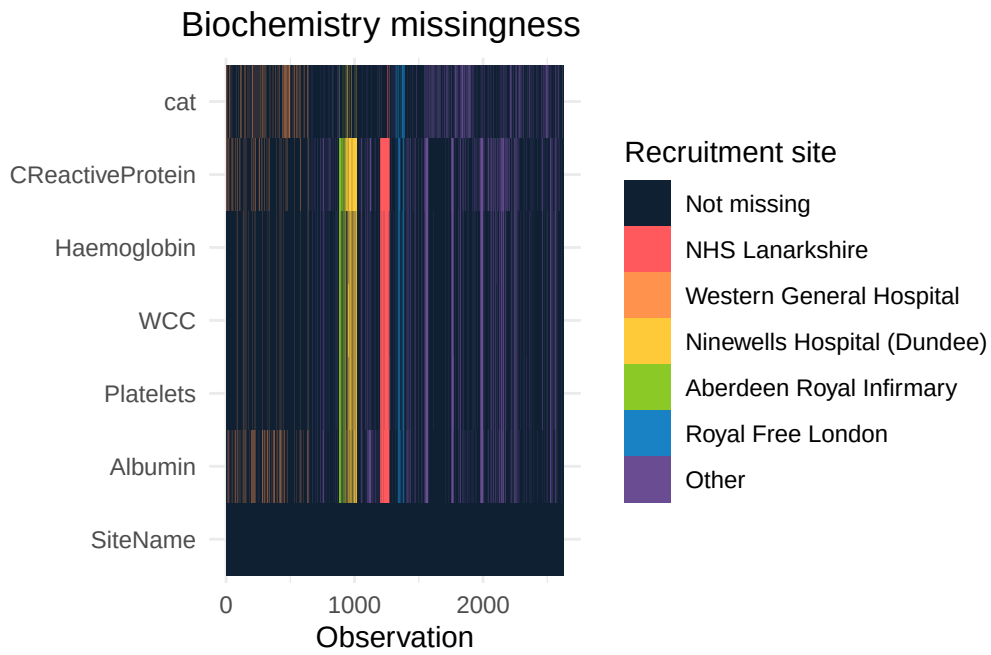
Table 6: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat	2	1069	534.5	23.85	6.098e-11
Residuals	1693	37943	22.41	NA	NA

Missingness

```
demo %>%
  select(
    cat,
    CReactiveProtein,
```

```
    Haemoglobin,
    WCC,
    Platelets,
    Albumin,
    SiteName
  ) %>%
missing_plot2(title = "Biochemistry missingness")
```




```
saveRDS(demo, paste0(outdir, "demo-biochem.RDS"))
```

Reproduction and reproducibility

Session info

R version 4.4.0 (2024-04-24)

Platform: aarch64-unknown-linux-gnu

locale: *LC_CTYPE=en_US.UTF-8*, *LC_NUMERIC=C*, *LC_TIME=en_US.UTF-8*, *LC_COLLATE=en_US.UTF-8*, *LC_MONETARY=en_US.UTF-8*, *LC_MESSAGES=en_US.UTF-8*, *LC_PAPER=en_US.UTF-8*, *LC_NAME=C*, *LC_ADDRESS=C*, *LC_TELEPHONE=C*, *LC_MEASUREMENT=en_US.UTF-8* and *LC_IDENTIFICATION=C*

attached base packages: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods* and *base*

other attached packages: *DiagrammeRsvg(v.0.1)*, *DiagrammeR(v.1.0.11)*, *pander(v.0.6.5)*, *knitr(v.1.47)*, *table1(v.1.4.3)*, *patchwork(v.1.2.0)*, *datefixR(v.1.6.1)*, *readxl(v.1.4.3)*, *lubridate(v.1.9.3)*, *forcats(v.1.0.0)*, *stringr(v.1.5.1)*, *dplyr(v.1.1.4)*, *purrr(v.1.0.2)*, *readr(v.2.1.5)*, *tidyr(v.1.3.1)*, *tibble(v.3.2.1)*, *ggplot2(v.3.5.1)*, *tidyverse(v.2.0.0)* and *plyr(v.1.8.9)*

loaded via a namespace (and not attached): *shape(v.1.4.6.1)*, *gtable(v.0.3.5)*, *xfun(v.0.44)*, *htmlwidgets(v.1.6.4)*, *visNetwork(v.2.1.2)*, *lattice(v.0.22-6)*, *tzdb(v.0.4.0)*, *vctrs(v.0.6.5)*, *tools(v.4.4.0)*, *generics(v.0.1.3)*, *curl(v.5.2.1)*, *fansi(v.1.0.6)*, *pan(v.1.9)*, *jomo(v.2.7-6)*, *pkgconfig(v.2.0.3)*, *Matrix(v.1.7-0)*, *RColorBrewer(v.1.1-3)*, *lifecycle(v.1.0.4)*, *compiler(v.4.4.0)*, *farver(v.2.1.2)*, *munsell(v.0.5.1)*, *codetools(v.0.2-20)*, *htmltools(v.0.5.8.1)*, *yaml(v.2.3.8)*, *finalfit(v.1.0.7)*, *glmnet(v.4.1-8)*, *Formula(v.1.2-5)*, *mice(v.3.16.0)*, *nloptr(v.2.0.3)*, *pillar(v.1.9.0)*, *MASS(v.7.3-60.2)*, *iterators(v.1.0.14)*, *rpart(v.4.1.23)*, *boot(v.1.3-30)*, *mitml(v.0.4-5)*, *foreach(v.1.5.2)*, *nlme(v.3.1-164)*, *tidyselect(v.1.2.1)*, *digest(v.0.6.35)*, *stringi(v.1.8.4)*, *splines(v.4.4.0)*, *labeling(v.0.4.3)*, *fastmap(v.1.2.0)*, *grid(v.4.4.0)*, *colorspace(v.2.1-0)*, *cli(v.3.6.2)*, *magrittr(v.2.0.3)*, *survival(v.3.5-8)*, *utf8(v.1.2.4)*, *broom(v.1.0.6)*, *withr(v.3.0.0)*, *scales(v.1.3.0)*, *backports(v.1.5.0)*, *timechange(v.0.3.0)*, *rmarkdown(v.2.27)*, *nnet(v.7.3-19)*, *lme4(v.1.1-35.3)*, *cellranger(v.1.1.0)*, *hms(v.1.1.3)*, *evaluate(v.0.23)*, *V8(v.4.4.2)*, *rlang(v.1.1.3)*, *Rcpp(v.1.0.12)*, *glue(v.1.7.0)*, *minqa(v.1.2.7)*, *jsonlite(v.1.8.8)* and *R6(v.2.5.1)*