Demographic data

Nathan Constantine-Cooke

2025-11-05

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 <- read_xlsx(paste0(data.path, "Baseline2022/demographics.xlsx"),
  col_types = c(
    "text",
    "text",
    "text",
    "text",
    "numeric",
    "numeric",
    "text",
    "text",
    "date",
    "numeric",
    "text"
  )
)

fcal <- read_xlsx(paste0(data.path, "Baseline2022/calprotectin.xlsx"))
fcal$Result <- as.numeric(plyr::mapvalues(fcal$Result, from = "<20", to = 20))

fcal <- fcal[, c("ParticipantNo", "Result")]
```

```r
fcal.eof <- read_xlsx(paste0(prefix, "EOF_fcal.xlsx"))

fcal.eof <- subset(fcal.eof, IsBaseline == 1)
fcal.eof <- subset(fcal.eof, FCALLevel != ".")
fcal.eof$FCALLevel <- as.numeric(fcal.eof$FCALLevel)
fcal.eof <- fcal.eof[, c("ParticipantNo", "FCALLevel")]
names(fcal.eof)[2] <- "Result"

fcal <- rbind(fcal, fcal.eof)
fcal <- distinct(fcal, ParticipantNo, .keep_all = TRUE)

fcal$cat <- 0
for (i in 1:nrow(fcal)) {
  if (fcal[i, "Result"] < 50) fcal[i, "cat"] <- 1
  if ((fcal[i, "Result"] >= 50) & (fcal[i, "Result"] <= 250)) fcal[i, "cat"] <- 2
  if (fcal[i, "Result"] > 250) fcal[i, "cat"] <- 3
}


demo <- merge(demo, fcal[, c("ParticipantNo", "Result", "cat")],
  by = "ParticipantNo",
  all.x = TRUE,
  all.y = FALSE
)

demo$cat <- factor(demo$cat,
  levels = c(1, 2, 3),
  labels = c("FC < 50", "FC 50-250", "FC > 250")
)

names(demo)[12] <- "FC" # Result ~> FC
```

This page describes key demographic data collected by the PREdiCCt study.

Age

The distribution of age for the study cohort is approximately normally distributed.

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

```
  geom_histogram(bins = 30, fill = "#64c7ce", color = "#3B9EA4") +
  geom_density() +
  theme_minimal() +
  xlab("Age (years)") +
  ylab("Frequency")
```

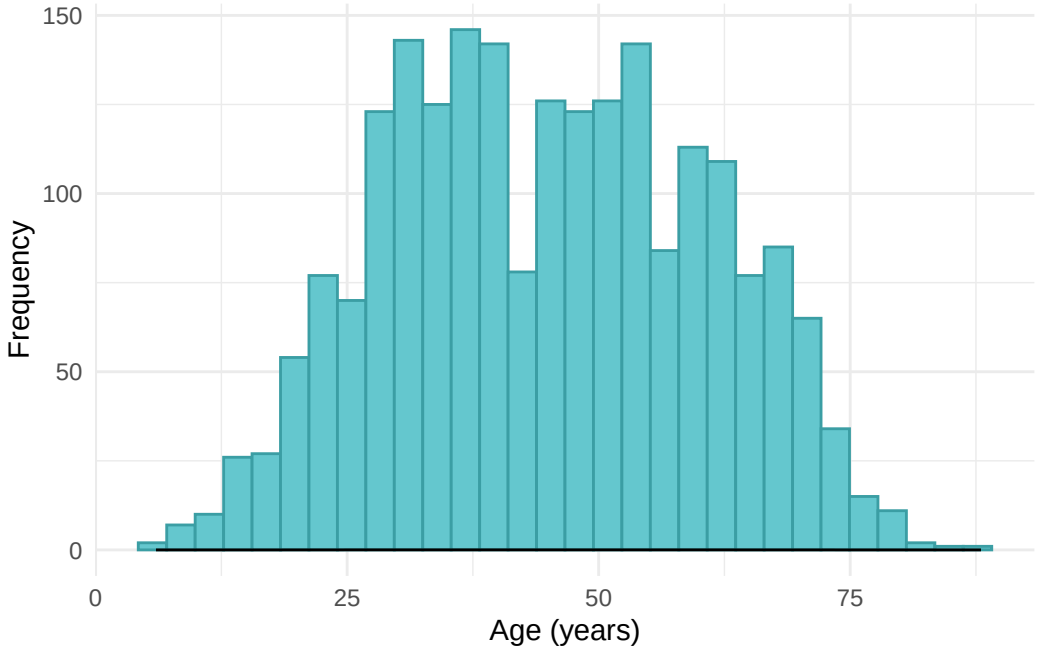


Figure 1: Distribution of age in the FC cohort.

Age is significantly associated with faecal calprotectin.

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

Table 1: ANOVA between age and FC groups.

Table 1: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat	2	2739	1370	5.661	0.003531
Residuals	2141	518041	242	NA	NA

Sex

Sex was self-reported via subject questionnaires. As can be seen in Figure 2, the PREdiCCt cohort is female by majority.

```
demo$Sex <- factor(demo$Sex, levels = c("1", "2"), labels = c("Male", "Female"))

demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = Sex, fill = Sex, color = Sex)) +
  geom_bar() +
  ylab("Frequency") +
  theme_minimal() +
  scale_fill_manual(values = c("#FCA17D", "#60D394")) +
  scale_color_manual(values = c("#B22A21", "#034C3C"))
```

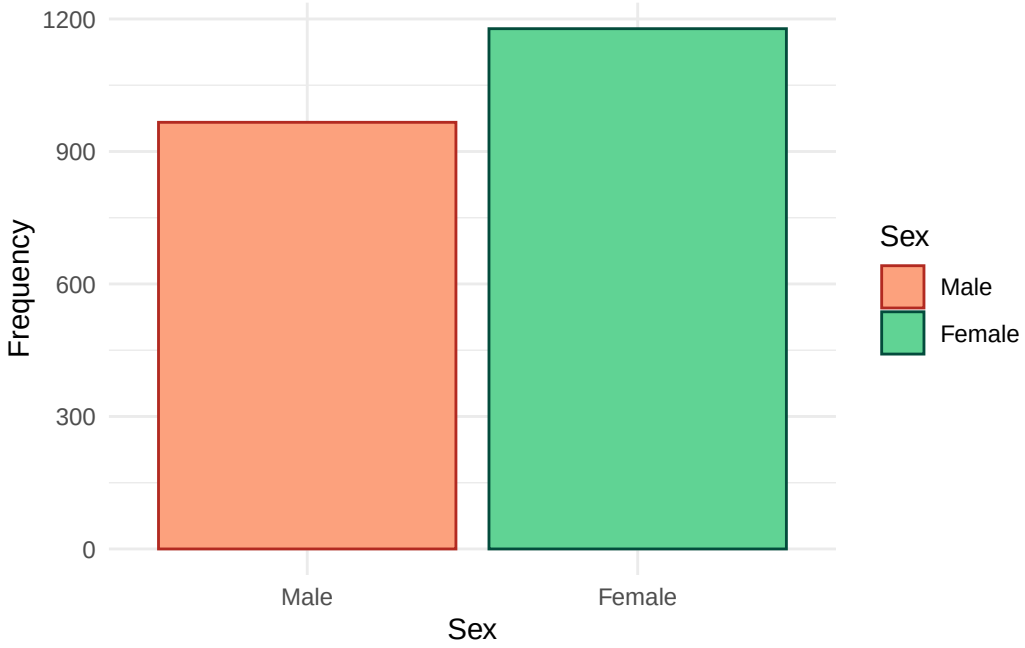


Figure 2: Distribution of sex in the FC cohort.

Moreover, the distribution of sex is significantly different between FCAL groups.

```
pander(chisq.test(demo$Sex, demo$cat))
```

Table 2: Chi-squared test between Sex and FC groups.

Table 2: Pearson's Chi-squared test: `demo$Sex` and `demo$cat`

Test statistic	df	P value
14.2	2	0.0008269 * * *

BMI

Body mass index is given by weight (in Kg) divided by height (in M) squared.

As thresholds differ in the paediatric population, and are dependent on both age and sex, BMI has not been calculated for those less than 18 years of age. BMI is first considered as a continuous variable.

```
demo$BMI <- with(demo, Weight / ((Height / 100)^2))
demo[demo[, "age"] < 18, "BMI"] <- NA
p <- demo %>%
  ggplot(aes(x = BMI)) +
  geom_histogram(color = "#AD013B", fill = "#F0386B", bins = 30) +
  theme_minimal(base_family = "sans") +
  ylab("Frequency")

ggsave("plots/BMI-full-cohort.png",
  p,
  width = 9,
  height = 6
)
ggsave("plots/BMI-full-cohort.pdf", p, width = 9, height = 6)


demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = BMI)) +
  geom_histogram(color = "#AD013B", fill = "#F0386B", bins = 30) +
  theme_minimal() +
  ylab("Frequency")
```

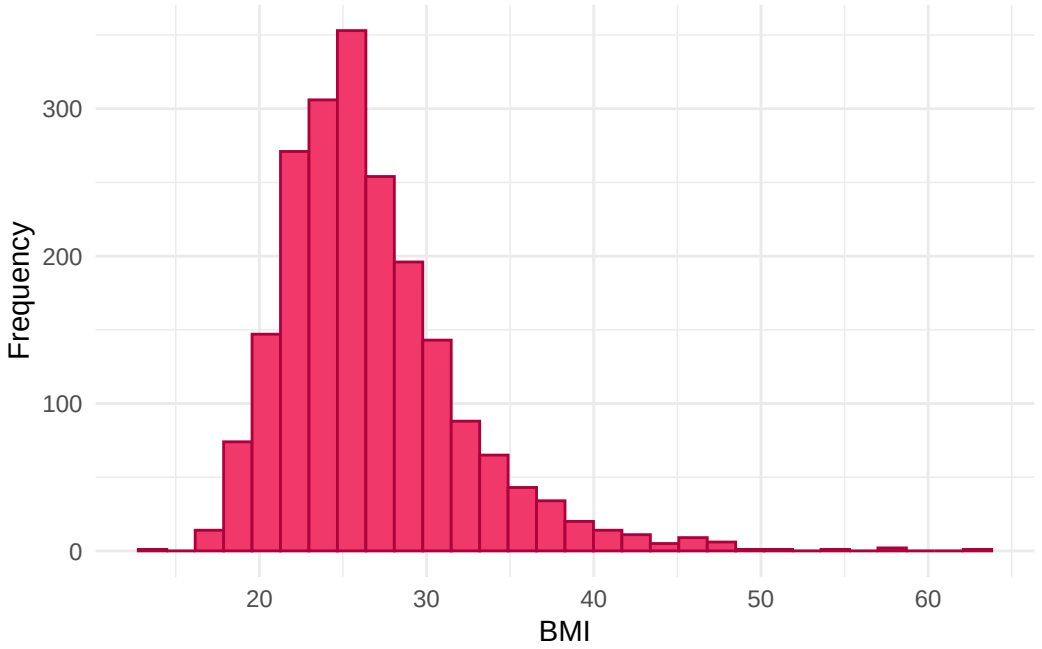


Figure 3: Distribution of BMI in the FC cohort.

BMI as a continuous variable is not associated with FC groups.

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

Table 3: Fisher's exact test between BMI and FC groups.

Table 3: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat	2	134.9	67.45	2.387	0.09219
Residuals	2057	58135	28.26	NA	NA

We also consider BMI grouped into underweight, normal, overweight, and obese categories using the definitions used by the NIH and WHO (Weir and Jan 2024).

A cut-off of $30 KgM^{-2}$ is used to denote obesity in adults without Asian/South Asian backgrounds. As data on Asian and South Asian ethnicity is not available (See Ethnicity), we are unable to adjust the threshold based on ethnicity. However, this is expected to be relevant to relatively few subjects.

```r
demo$BMIcat <- cut(demo$BMI,
  c(0, 18.5, 25, 30, Inf),
  include.lowest = TRUE,
  right = FALSE,
  labels = c("Underweight", "Normal", "Overweight", "Obese")
)
demo[demo[, "age"] < 18, "BMIcat"] <- NA

plt.cols <- c("#F6AE2D", "#C81D25", "#1481BA", "#7E1F86")


demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = BMIcat, color = BMIcat, fill = BMIcat)) +
  geom_bar() +
  theme_minimal() +
  theme(legend.position = "none") +
  scale_fill_manual(values = plt.cols, na.value = "#032B43") +
  scale_color_manual(
    values = colorspace::darken(plt.cols, 0.2),
    na.value = "#032B43"
  ) +
  xlab("BMI category") +
  ylab("Frequency")
```

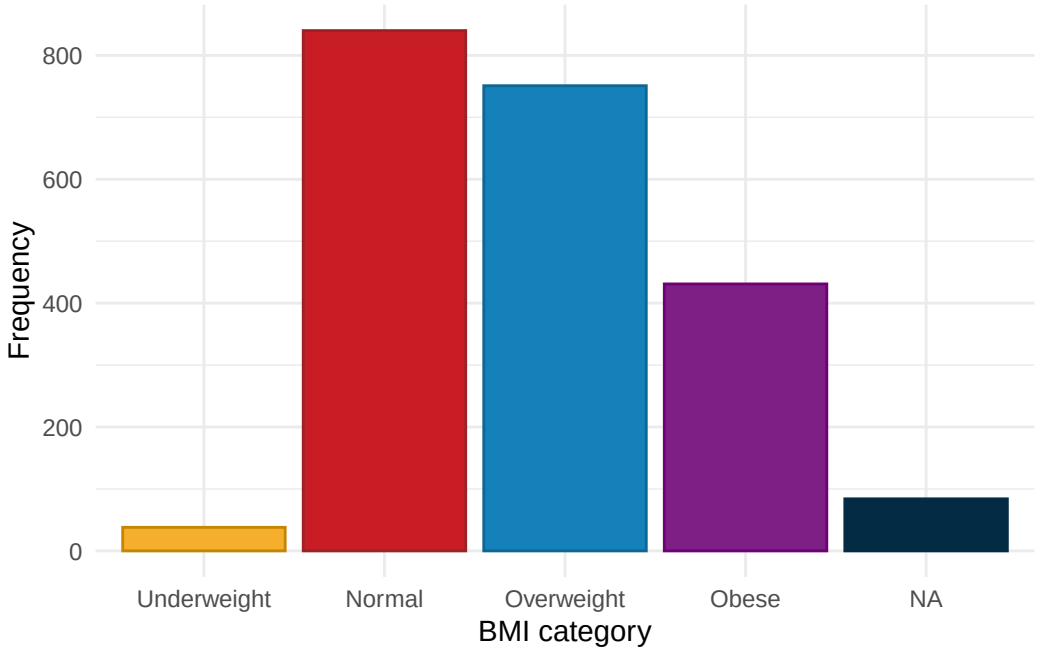


Figure 4: Distribution of BMI categories in the FC cohort.

BMI by category is not associated with FC groups.

```
pander(chisq.test(demo$BMIcat, demo$cat))
```

Table 4: Chi-squared test between BMI categories and FC groups.

Table 4: Pearson's Chi-squared test: `demo$BMIcat` and `demo$cat`

Test statistic	df	P value
7.169	6	0.3055

Ethnicity

Due to low counts for non-white ethnicities, we are only able to report frequencies of white and non-white ethnicities.

```
demo$ethnic_gp <- factor(demo$ethnic_gp,
  levels = c("1", "2"),
  labels = c("White", "Non-white")
```

```
)
colnames(demo)[10:11] <- c("Age", "Ethnicity")


demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = Ethnicity, fill = Ethnicity, color = Ethnicity)) +
  geom_bar() +
  ylab("Frequency") +
  theme_minimal() +
  scale_fill_manual(values = c("#F24236", "#2E86AB"), na.value = "#032B43") +
  scale_color_manual(values = c("#C0362C", "#1E556C"), na.value = "#032B43")
```

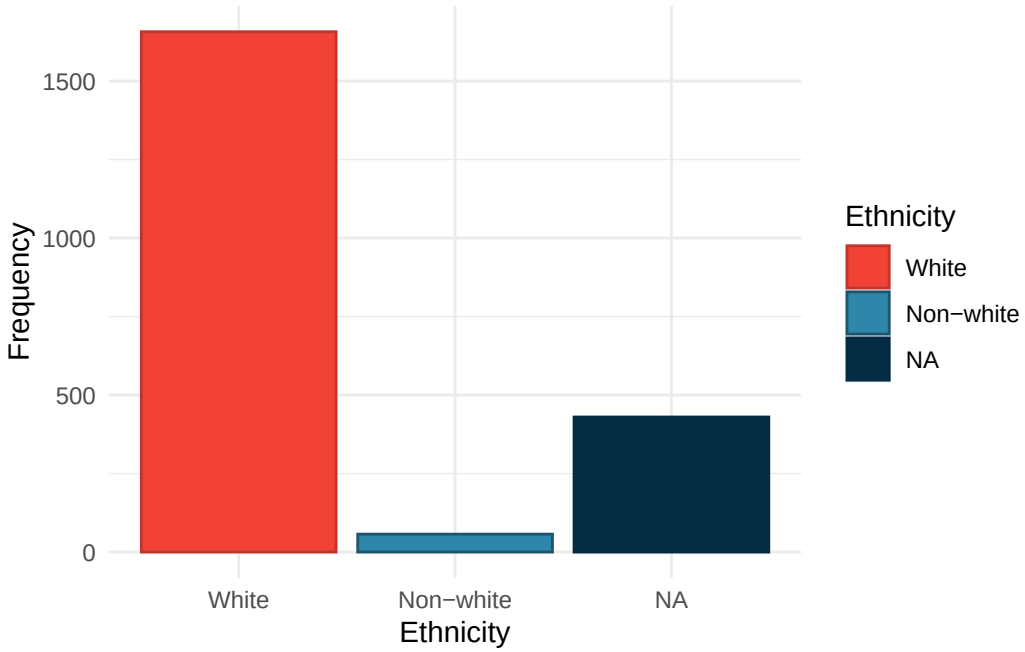


Figure 5: Distribution of ethnicity in the FC cohort.

```
pander(chisq.test(demo$Ethnicity, demo$cat))
```

Table 5: Chi-squared test between ethnicity and FC groups.

Table 5: Pearson's Chi-squared test: `demo$Ethnicity` and `demo$cat`

Test statistic	df	P value
2.416	2	0.2989

Index of multiple deprivation

The PREdiCCt SAP states the primary analyses will be controlled for by index of multiple deprivation (IMD). It should be noted that the PREdiCCt study recruited across the UK and there is no consistent measure of IMD across the whole of the UK. Instead, IMD measures are handled slightly differently by each constituent nation.

```
IMD <- read_xlsx(paste0(redcap.path, "/IMD.xlsx"))


IMD <- as.data.frame(IMD)
names(IMD)[2] <- "IMD"
demo <- merge(demo, IMD, by = "ParticipantNo", all.x = TRUE, all.y = FALSE)
demo$IMD <- as.factor(demo$IMD)

cols <- c("#A8F9FF", "#EF8275", "#6D326D", "#0B6E4F", "#F4D35E")

demo %>%
  drop_na(cat) %>%
  ggplot(aes(x = IMD, color = as.factor(IMD), fill = as.factor(IMD))) +
  geom_bar() +
  theme_minimal() +
  theme(legend.position = "none") +
  ylab("Frequency") +
  xlab("IMD (most deprived to least deprived)") +
  scale_fill_manual(values = cols) +
  scale_color_manual(values = colorspace::darken(cols, amount = 0.3))
```

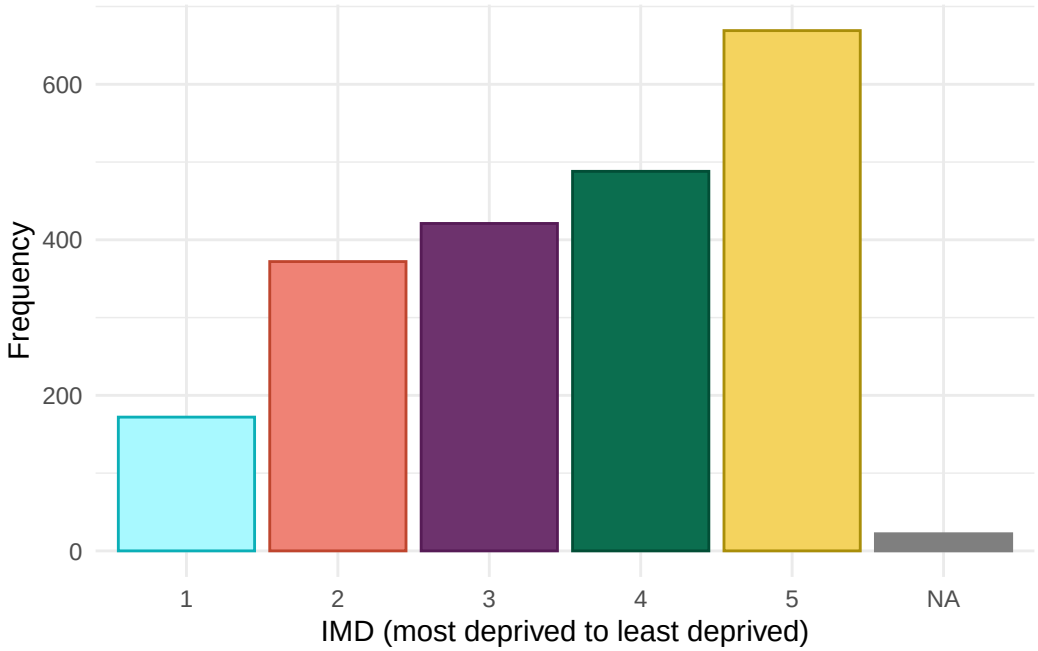


Figure 6: Distribution of IMD.

```
pander(chisq.test(demo$IMD, demo$cat))
```

Table 6: CHI-squared test between IMD and FC groups.

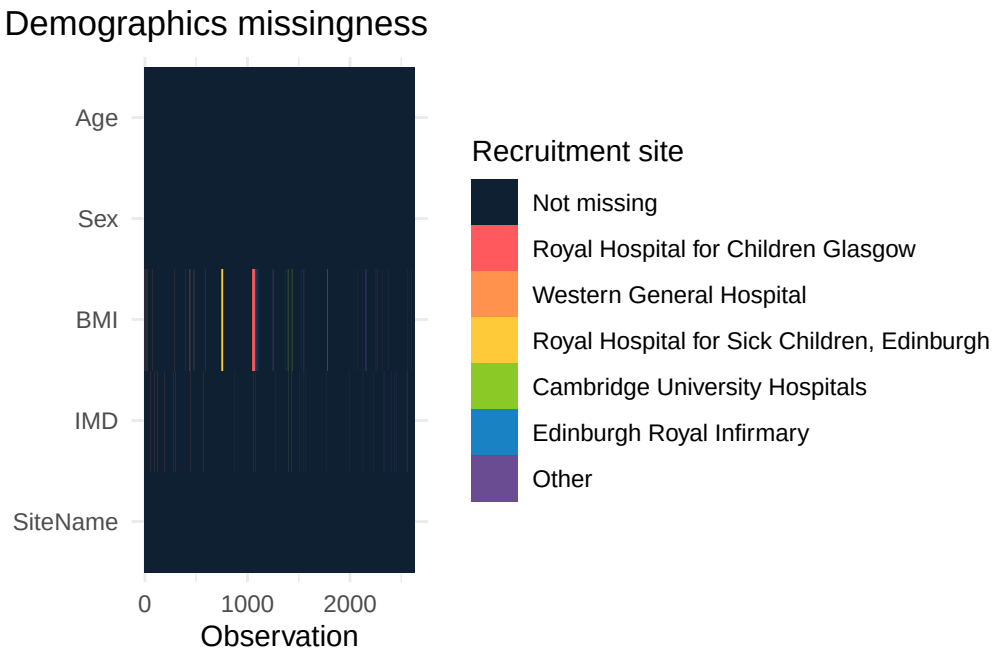
Table 6: Pearson's Chi-squared test: `demo$IMD` and `demo$cat`

Test statistic	df	P value
8.464	8	0.3895

Missingness

```
demo %>%
  select(
    Age,
    Sex,
    BMI,
    IMD,
    SiteName
```

```
  ) %>%
missing_plot2(title = "Demographics missingness")
```




```
saveRDS(demo, paste0(outdir, "demo-demographics.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)*, *textshaping(v.0.4.0)*, *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)*, *ragg(v.1.3.2)*, *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)*, *R6(v.2.5.1)* and *systemfonts(v.1.3.1)*

Weir, Connor B., and Arif Jan. 2024. "BMI Classification Percentile and Cut Off Points." In *StatPearls*. Treasure Island (FL): StatPearls Publishing.