Study recruitment

Nathan Constantine-Cooke

2025-11-05

Table of contents


```
set.seed(123)
suppressPackageStartupMessages(library(tidyverse)) # ggplot2, dplyr, and magrittr
library(readxl) # Read in Excel files
# Generate flowchart of cohort derivation
library(DiagrammeR)
library(DiagrammeRsvg)

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/"
  metadir <- "data/metadata/"
} 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/"
  metadir <- "/Volumes/igmm/cvallejo-predicct/predicct/metadata/"
}
```

```r
eof <- read_xlsx(paste0(prefix, "Followup form.xlsx"))


# Demographic data as reported by subjects
demo <- read_xlsx(paste0(data.path, "Baseline2022/demographics.xlsx"),
  col_types = c(
    "text",
    "text",
    "text",
    "text",
    "numeric",
    "numeric",
    "text",
    "text",
    "date",
    "numeric",
    "text"
  )
)
```

Country of recruitment

PREdiCCt is a pan-UK study which recruited across 47 sites. Figure 1 shows the distribution of the PREdiCCt cohort by country of the recruiting site.

```r
site.data <- read.csv(paste0(metadir, "sites.csv"))
sites <- demo %>%
  select(ParticipantNo, SiteNo)
sites <- merge(sites, site.data, by = "SiteNo", all.x = TRUE, all.y = FALSE)

plt.cols <- c( "#003459", "#DB5461", "#41D3BD", "#F09D51")

sites$Country <- factor(sites$Country,
                        levels = c("Scotland",
                                   "England",
                                   "Wales",
                                   "Northern Ireland"))


sites %>%
  ggplot(aes(x = Country, color = Country, fill = Country)) +
```

```
  geom_bar() +
  theme_minimal() +
  theme(legend.position = "none") +
  xlab("Country of recruiting site") +
  ylab("Frequency") +
  scale_fill_manual(values = plt.cols) +
  scale_color_manual(values = colorspace::darken(plt.cols, 0.2))
```

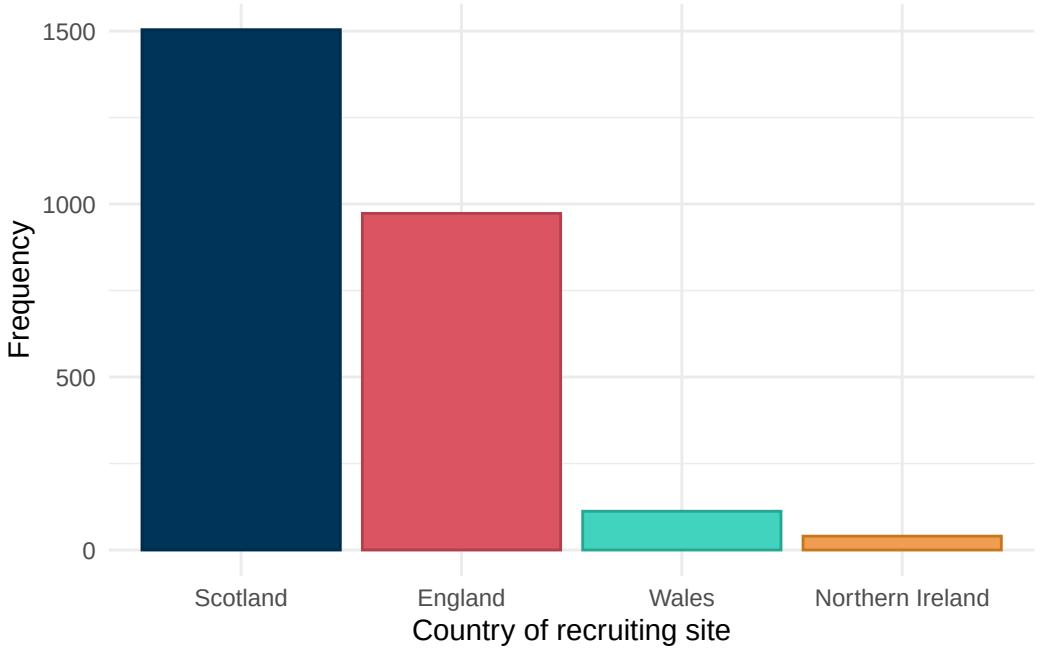

Figure 1: Distribution of the PREdiCCt cohort by country of recruitment site.

Recruitment over time

```
demo <- demo %>%  mutate(entry_date = as.Date(entry_date))
```

Recruitment to PREdiCCt began in November 2016. As participants were recruited when they attended IBD clinic appointments and the COVID-19 pandemic substantially decreased the number of in-person clinic appointments, recruitment was ceased in March 2020.

```r
demo_cumulative <- demo %>%
  arrange(entry_date) %>%
  mutate(cumulative_count = row_number())

p <- demo_cumulative %>%
  ggplot(aes(x = entry_date, y = cumulative_count)) +
  geom_smooth(color = rgb(34, 122, 145, maxColorValue = 255),
              method = "gam") +
  theme_minimal() +
  xlab("Year") +
  ylab("Study recruitment") +
  xlim(as.Date("2016-11-01"), as.Date("2020-04-01")) +
  scale_y_continuous(breaks = seq(0, 3000, by = 500), limits = c(0, 3000)) +
  theme(text = element_text(color = "#1C285A"),
        axis.text = element_text(face = "bold", color = "#1C285A"))

#ggsave("src/plots/baseline/cumulative-recruitment.png",
#       p,
#       width = 9 * 0.8,
#       height = 6 * 0.8)

#ggsave("src/plots/baseline/cumulative-recruitment.pdf",
#       p,
#       width = 9 * 0.8,
#       height = 6 * 0.8)
p
```

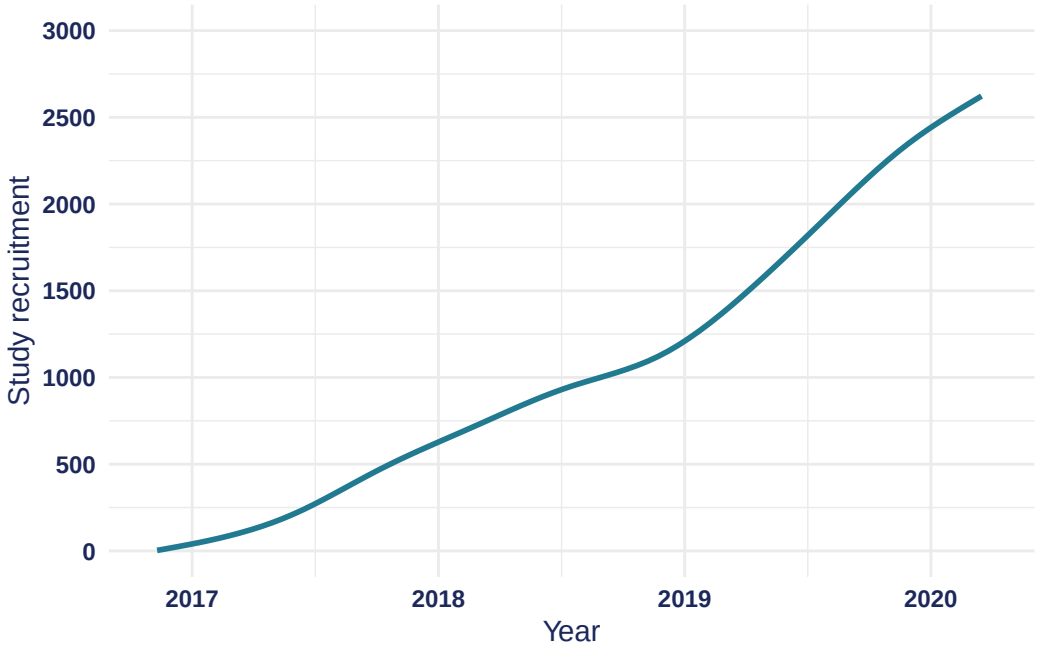

Figure 2: Cumulative recruitment to the PREdiCCt study over time.

Month of recruitment

There is potential for seasonality to confound interpretations of our results, particularly when making inferences regarding diet. As such, month of recruitment is explored (Figure 3).

Recruitment was low in December, which is to be as expected as there are typically fewer clinic appointments in December and participants were recruited in IBD clinics.

However, recruitment was also low in April and July. As such, the impact of seasonality is likely to be minimal for this cohort and will be ignored for all analyses.

```
demo$entry_date <- as.Date(demo$entry_date)
demo$month <- month(demo$entry_date, label = TRUE)
ggplot(demo, aes(x = as.factor(month), color = as.factor(month), fill = as.factor(month))) +
  geom_bar() +
  theme_minimal() +
  theme(legend.position = "none") +
  xlab("Month of recruitment") +
  ylab("Frequency")
```

Figure 3: Distribution of month of diagnosis.

Cohort derivation

The below flowchart gives a simple explanation of how the sub-cohort of subjects with analysed food frequency questionnaires (FFQs) was obtained.

```
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")]

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)

demo <- merge(demo, fcal, by = "ParticipantNo", all.x = TRUE, all.y = FALSE)

FFQ <- read_xlsx(paste0(
  prefix,
  "predicct ffq_nutrientfood groupDQI all foods_data (n1092)Nov2022.xlsx"
))

FFQ <- subset(FFQ, participantno %in% fcal$ParticipantNo)


flow <- grViz(paste0("digraph flowchart {

graph[splines = ortho]
  # node definitions with substituted label text
  node [fontname = Helvetica, shape = rectangle, fixedsize = false, width = 1]
  1 [label = 'Recruited to PREdiCCt\n n = ", nrow(demo), "']
  2 [label = 'Baseline FC available\n n = ", nrow(drop_na(demo, Result)), "']
  3 [label = 'Food frequency questionnaire\n analysed\n n = ", nrow(FFQ), "']

  node [shape=none, width=0, height=0, label='']
  1 -> 2; 2-> 3
}"))
htmltools::HTML(export_svg(flow))
```

End of study phenotyping

```
eof <- subset(eof, DiseaseFlareYN != ".")
```

End of study phenotyping was avaialable for 2476 subjects. This equates to 94.2% of the total cohort.

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)*, *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)* and *tidyverse(v.2.0.0)*

loaded via a namespace (and not attached): *utf8(v.1.2.4)*, *generics(v.0.1.3)*, *lattice(v.0.22-6)*, *stringi(v.1.8.4)*, *hms(v.1.1.3)*, *digest(v.0.6.35)*, *magrittr(v.2.0.3)*, *evaluate(v.0.23)*, *grid(v.4.4.0)*, *timechange(v.0.3.0)*, *RColorBrewer(v.1.1-3)*, *fastmap(v.1.2.0)*, *plyr(v.1.8.9)*, *Matrix(v.1.7-0)*, *cellranger(v.1.1.0)*, *jsonlite(v.1.8.8)*, *mgcv(v.1.9-1)*, *pander(v.0.6.5)*, *fansi(v.1.0.6)*, *viridisLite(v.0.4.2)*, *scales(v.1.3.0)*, *codetools(v.0.2-20)*, *cli(v.3.6.2)*, *rlang(v.1.1.3)*, *visNetwork(v.2.1.2)*, *splines(v.4.4.0)*, *munsell(v.0.5.1)*, *withr(v.3.0.0)*, *yaml(v.2.3.8)*, *tools(v.4.4.0)*, *tzdb(v.0.4.0)*, *colorspace(v.2.1-0)*, *curl(v.5.2.1)*, *vctrs(v.0.6.5)*, *R6(v.2.5.1)*, *lifecycle(v.1.0.4)*, *V8(v.4.4.2)*, *htmlwidgets(v.1.6.4)*, *pkgconfig(v.2.0.3)*, *pillar(v.1.9.0)*, *gtable(v.0.3.5)*, *glue(v.1.7.0)*, *Rcpp(v.1.0.12)*, *xfun(v.0.44)*, *tidyselect(v.1.2.1)*, *rstudioapi(v.0.16.0)*, *knitr(v.1.47)*, *farver(v.2.1.2)*, *nlme(v.3.1-164)*, *htmltools(v.0.5.8.1)*, *labeling(v.0.4.3)*, *rmarkdown(v.2.27)* and *compiler(v.4.4.0)*