

Overall tables

Nathan Constantine-Cooke

2025-08-13

Table of contents

Introduction	2
Table of baseline data by cohort	3
Associations with cohort membership	7
Age	7
Sex	7
Body mass index	7
Ethnicity	8
Index of multiple deprivation	8
Smoking status	8
IBD type	8
Disease duration	9
IBD Control-8	9
IBD visual analogue score	9
Faecal calprotectin	9
C-reactive protein	10
Haemoglobin	10
White cell count	10
Platelets	11
Albumin	11
Dietary data	11
Biologic usage	11
Table of baseline data by FC groups	12
Table of baseline data by IBD type	12
Table of Crohn's disease variables	14
Table of Crohn's disease variables by cohort	18
Associations between Crohn's Disease-only variables and cohort membership	20
Montreal location	20
Upper gastrointestinal inflammation	20

Harvey-Bradshaw index	20
PRO2	20
Montreal behaviour	21
Perianal disease	21
Surgery	21
Smoking status	22
E-cigarette use	22
Table of ulcerative colitis/IBDU variables	22
Table of Ulcerative colitis/IBDU variables by cohort	23
Associations between UC/IBDU-only variables and cohort membership	25
Montreal extent	25
Mayo score	26
PRO2	26
Smoking status	26
E-cigarette use	26
Reproduction and reproducibility	27

Introduction

```

set.seed(123)

#####
## 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

# 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-diet.RDS"))
FFQ <- read_xlsx(paste0(
  prefix,
  "predicct_ffq_nutrientfood_groupDQI_all_foods_data (n1092)Nov2022.xlsx"
))

```

On this page, you will find key demographic/phenotypic tables for the cohorts. You will also find statistical tests exploring if participant characteristics differ across cohorts.

In addition to the FC cohort (PREDiCCt subjects with a baseline FC available), there is also the FFQ cohort which consists of subjects with analysed FFQs available. All subjects in the FFQ cohort also have a baseline FC available and are therefore also in the FC cohort.

Table of baseline data by cohort

```

my.render.cont <- function(x) {
  with(
    stats.apply.rounding(stats.default(x),
      digits = 3,
      round.integers = FALSE
    ),
    c("", "Median (IQR)" = sprintf("%s (%s - %s)", MEDIAN, Q1, Q3))
  )
}

demo$control_8 <- as.numeric(demo$control_8)

```

```

comp <- demo
comp$cohort <- "All"

temp <- demo %>%
  drop_na(cat)
temp$cohort <- "FC"

comp <- rbind(comp, temp)

temp <- subset(demo, ParticipantNo %in% FFQ$participantno)
temp$cohort <- "FFQ"

comp <- rbind(comp, temp)

comp$cohort <- factor(comp$cohort,
  levels = c("All", "FC", "FFQ"),
  labels = c("Full cohort", "FC cohort", "FFQ cohort")
)

table1(
  ~ Age +
  Sex +
  Ethnicity +
  BMICat +
  diagnosis +
  `IBD Duration` +
  as.numeric(IMD) +
  Smoke +
  ECigs +
  control_8 +
  vas_control +
  FC +
  CReactiveProtein +
  Haemoglobin +
  WCC +
  Albumin +
  Meat_sum +
  fibre +
  PUFA_percEng +
  NOVAScore_cat +
  Biologic | cohort,
  data = comp,

```

```
render.continuous = my.render.cont,  
overall = FALSE  
)
```

	Full cohort	FC cohort	FFQ cohort
	(N=2629)	(N=2144)	(N=1091)
Age			
Median (IQR)	44.0 (32.0 - 56.0)	44.0 (32.0 - 57.0)	47.0 (35.0 - 58.0)
Sex			
Male	1207 (45.9%)	966 (45.1%)	476 (43.6%)
Female	1422 (54.1%)	1178 (54.9%)	615 (56.4%)
Ethnicity			
White	1862 (70.8%)	1657 (77.3%)	1039 (95.2%)
Non-white	71 (2.7%)	57 (2.7%)	24 (2.2%)
Missing	696 (26.5%)	430 (20.1%)	28 (2.6%)
BMIcat			
Underweight	49 (1.9%)	38 (1.8%)	20 (1.8%)
Normal	1004 (38.2%)	840 (39.2%)	460 (42.2%)
Overweight	923 (35.1%)	751 (35.0%)	398 (36.5%)
Obese	552 (21.0%)	431 (20.1%)	191 (17.5%)
Missing	101 (3.8%)	84 (3.9%)	22 (2.0%)
diagnosis			
CD	1370 (52.1%)	1118 (52.1%)	530 (48.6%)
UC	1174 (44.7%)	950 (44.3%)	523 (47.9%)
IBDU	85 (3.2%)	76 (3.5%)	38 (3.5%)
IBD Duration			
Median (IQR)	10.0 (4.73 - 18.6)	9.80 (4.60 - 18.6)	10.3 (4.92 - 20.1)
Missing	151 (5.7%)	120 (5.6%)	53 (4.9%)
as.numeric(IMD)			
Median (IQR)	4.00 (2.00 - 5.00)	4.00 (2.00 - 5.00)	4.00 (3.00 - 5.00)
Missing	30 (1.1%)	22 (1.0%)	9 (0.8%)
Smoke			
Current	122 (4.6%)	109 (5.1%)	63 (5.8%)
Previous	721 (27.4%)	627 (29.2%)	399 (36.6%)
Never	1071 (40.7%)	965 (45.0%)	597 (54.7%)
Missing	715 (27.2%)	443 (20.7%)	32 (2.9%)
ECigs			
Current	80 (3.0%)	70 (3.3%)	38 (3.5%)
Previous	61 (2.3%)	53 (2.5%)	27 (2.5%)
Never	1770 (67.3%)	1575 (73.5%)	994 (91.1%)
Missing	718 (27.3%)	446 (20.8%)	32 (2.9%)
control_8			
Median (IQR)	13.0 (11.0 - 15.0)	13.0 (11.0 - 15.0)	13.0 (11.0 - 15.0)
Missing	718 (27.3%)	447 (20.8%)	38 (3.5%)
vas_control			
<85	628 (23.9%)	569 (26.5%)	320 (29.3%)
85+	1275 (48.5%)	1121 (52.3%)	730 (66.9%)
Missing	726 (27.6%)	454 (21.2%)	41 (3.8%)
FC			
Median (IQR)	49.0 (20.0 - 161)	49.0 (20.0 - 161)	49.0 (20.0 - 161)
Missing	485 (18.4%)	0 (0%)	79 (7.2%)
CReactiveProtein			
Median (IQR)	2.00 (1.00 - 5.00)	2.00 (1.00 - 5.00)	2.00 (1.00 - 5.00)

Associations with cohort membership

Only age, biologic use, and albumin were found to be significantly different across cohorts.

Age

There is a significant difference in age across cohorts with subjects in the FFQ sub-cohort being more likely to be older than the full or FC cohorts.

```
pander(summary(aov(Age ~ cohort, data = comp)))
```

Table 1: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	4786	2393	10.08	4.256e-05
Residuals	5861	1391260	237.4	NA	NA

Sex

```
pander(chisq.test(comp$Sex, comp$cohort))
```

Table 2: Pearson's Chi-squared test: `comp$Sex` and `comp$cohort`

Test statistic	df	P value
1.639	2	0.4406

Body mass index

```
pander(summary(aov(BMI ~ cohort, data = comp)))
```

Table 3: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	111.4	55.7	1.95	0.1424
Residuals	5654	161505	28.56	NA	NA

Ethnicity

```
pander(chisq.test(comp$Ethnicity, comp$cohort))
```

Table 4: Pearson's Chi-squared test: `comp$Ethnicity` and `comp$cohort`

Test statistic	df	P value
4.482	2	0.1063

Index of multiple deprivation

```
pander(chisq.test(comp$IMD, comp$cohort))
```

Table 5: Pearson's Chi-squared test: `comp$IMD` and `comp$cohort`

Test statistic	df	P value
10.73	8	0.2172

Smoking status

```
pander(chisq.test(comp$Smoke, comp$cohort))
```

Table 6: Pearson's Chi-squared test: `comp$Smoke` and `comp$cohort`

Test statistic	df	P value
0.5422	4	0.9693

IBD type

```
pander(chisq.test(comp$diagnosis2, comp$cohort))
```

Table 7: Pearson's Chi-squared test: `comp$diagnosis2` and `comp$cohort`

Test statistic	df	P value
4.474	2	0.1068

Disease duration

```
pander(summary(aov(`IBD Duration` ~ cohort, data = comp)))
```

Table 8: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	356.2	178.1	1.492	0.225
Residuals	5537	660933	119.4	NA	NA

IBD Control-8

```
pander(summary(aov(control_8 ~ cohort, data = comp)))
```

Table 9: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	19.77	9.884	1.018	0.3613
Residuals	4658	45211	9.706	NA	NA

IBD visual analogue score

```
pander(chisq.test(comp$vas_control, comp$cohort))
```

Table 10: Pearson's Chi-squared test: `comp$vas_control` and `comp$cohort`

Test statistic	df	P value
3.158	2	0.2062

Faecal calprotectin

FC has been treated as a continuous variable for this test. FC has not been discretised.

```
pander(summary(aov(FC ~ cohort, data = comp)))
```

Table 11: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	10042	5021	0.04235	0.9585
Residuals	5297	6.28e+08	118564	NA	NA

C-reactive protein

```
pander(summary(aov(CReactiveProtein ~ cohort, data = comp)))
```

Table 12: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	12.36	6.178	0.05603	0.9455
Residuals	4438	489297	110.3	NA	NA

Haemoglobin

```
pander(summary(aov(Haemoglobin ~ cohort, data = comp)))
```

Table 13: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	84.22	42.11	0.2276	0.7965
Residuals	4834	894479	185	NA	NA

White cell count

```
pander(summary(aov(WCC ~ cohort, data = comp)))
```

Table 14: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	81.86	40.93	1.06	0.3464
Residuals	4831	186499	38.6	NA	NA

Platelets

```
pander(summary(aov(Platelets ~ cohort, data = comp)))
```

Table 15: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	12234	6117	1.25	0.2866
Residuals	4830	23635277	4893	NA	NA

Albumin

There is a significant difference in albumin across the cohorts. However, this difference appears to be negligible.

```
pander(summary(aov(Albumin ~ cohort, data = comp)))
```

Table 16: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	320.6	160.3	6.931	0.0009876
Residuals	4591	106189	23.13	NA	NA

Dietary data

Dietary data have not been compared across cohorts.

Biologic usage

There is a significant difference in biologic usage across the cohorts with subjects in the FFQ sub-cohort being less likely to have been prescribed a biologic than the full cohort or the FC sub-cohort.

```
pander(chisq.test(comp$Biologic, comp$cohort))
```

Table 17: Pearson's Chi-squared test: `comp$Biologic` and `comp$cohort`

Test statistic	df	P value
21.41	4	0.0002626 * * *

Table of baseline data by FC groups

```
demo %>%
  drop_na(cat) %>%
  table1(
    x = ~ Age +
    Sex +
    Ethnicity +
    BMICat +
    diagnosis +
    `IBD Duration` +
    Smoke +
    ECigs +
    as.numeric(IMD) +
    control_8 +
    vas_control +
    FC +
    CReactiveProtein +
    Haemoglobin +
    WCC +
    Platelets +
    Albumin +
    Meat_sum +
    fibre +
    PUFA_percEng +
    NOVAScore_cat +
    Biologic | cat,
    render.continuous = my.render.cont
  )
```

Table of baseline data by IBD type

Table 18: Baseline data by FC.

	FC < 50	FC 50-250	FC > 250	Overall
	(N=1077)	(N=683)	(N=384)	(N=2144)
Age				
Median (IQR)	44.0 (33.0 - 56.0)	46.0 (33.0 - 59.5)	42.5 (30.0 - 55.0)	44.0 (32.0 - 57.0)
Sex				
Male	442 (41.0%)	333 (48.8%)	191 (49.7%)	966 (45.1%)
Female	635 (59.0%)	350 (51.2%)	193 (50.3%)	1178 (54.9%)
Ethnicity				
White	834 (77.4%)	540 (79.1%)	283 (73.7%)	1657 (77.3%)
Non-white	33 (3.1%)	13 (1.9%)	11 (2.9%)	57 (2.7%)
Missing	210 (19.5%)	130 (19.0%)	90 (23.4%)	430 (20.1%)
BMIcat				
Underweight	19 (1.8%)	12 (1.8%)	7 (1.8%)	38 (1.8%)
Normal	446 (41.4%)	244 (35.7%)	150 (39.1%)	840 (39.2%)
Overweight	371 (34.4%)	246 (36.0%)	134 (34.9%)	751 (35.0%)
Obese	201 (18.7%)	155 (22.7%)	75 (19.5%)	431 (20.1%)
Missing	40 (3.7%)	26 (3.8%)	18 (4.7%)	84 (3.9%)
diagnosis				
CD	533 (49.5%)	381 (55.8%)	204 (53.1%)	1118 (52.1%)
UC	504 (46.8%)	283 (41.4%)	163 (42.4%)	950 (44.3%)
IBDU	40 (3.7%)	19 (2.8%)	17 (4.4%)	76 (3.5%)
IBD Duration				
Median (IQR)	10.1 (4.62 - 19.4)	9.66 (4.76 - 18.4)	9.07 (4.42 - 16.5)	9.80 (4.60 - 18.6)
Missing	60 (5.6%)	38 (5.6%)	22 (5.7%)	120 (5.6%)
Smoke				
Current	64 (5.9%)	36 (5.3%)	9 (2.3%)	109 (5.1%)
Previous	309 (28.7%)	213 (31.2%)	105 (27.3%)	627 (29.2%)
Never	487 (45.2%)	300 (43.9%)	178 (46.4%)	965 (45.0%)
Missing	217 (20.1%)	134 (19.6%)	92 (24.0%)	443 (20.7%)
ECigs				
Current	28 (2.6%)	33 (4.8%)	9 (2.3%)	70 (3.3%)
Previous	31 (2.9%)	16 (2.3%)	6 (1.6%)	53 (2.5%)
Never	799 (74.2%)	500 (73.2%)	276 (71.9%)	1575 (73.5%)
Missing	219 (20.3%)	134 (19.6%)	93 (24.2%)	446 (20.8%)
as.numeric(IMD)				
Median (IQR)	4.00 (3.00 - 5.00)	4.00 (2.00 - 5.00)	4.00 (2.00 - 5.00)	4.00 (2.00 - 5.00)
Missing	15 (1.4%)	4 (0.6%)	3 (0.8%)	22 (1.0%)
control_8				
Median (IQR)	13.0 (11.0 - 15.0)	13.0 (11.0 - 15.0)	13.0 (9.00 - 15.0)	13.0 (11.0 - 15.0)
Missing	218 (20.2%)	136 (19.9%)	93 (24.2%)	447 (20.8%)
vas_control				
<85	250 (23.2%)	201 ¹³ (29.4%)	118 (30.7%)	569 (26.5%)
85+	607 (56.4%)	342 (50.1%)	172 (44.8%)	1121 (52.3%)
Missing	220 (20.4%)	140 (20.5%)	94 (24.5%)	454 (21.2%)
FC				
Median (IQR)	20.0 (20.0 - 29.0)	104 (71.0 - 157)	498 (347 - 780)	49.0 (20.0 - 161)
CReactiveProtein				
Median (IQR)	2.00 (1.00 - 5.00)	2.00 (1.00 - 5.00)	4.00 (1.00 - 6.00)	2.00 (1.00 - 5.00)

```

demo %>%
  drop_na(cat) %>%
  table1::table1(
    x = ~ Age +
    Sex +
    Ethnicity +
    BMICat +
    diagnosis +
    `IBD Duration` +
    Smoke +
    ECigs +
    as.numeric(IMD) +
    control_8 +
    vas_control +
    FC +
    CReactiveProtein +
    Haemoglobin +
    WCC +
    Platelets +
    Albumin +
    Meat_sum +
    fibre +
    PUFA_percEng +
    NOVAScore_cat +
    Biologic |
      diagnosis2,
    render.continuous = my.render.cont
  )

```

Table of Crohn's disease variables

```

demo.cd <- readRDS(paste0(outdir, "demo-cd.RDS"))
demo.cd %>%
  drop_na(cat) %>%
  table1(
    x = ~ `IBD Duration` +
    Location +
    L4 +
    HBI +

```

Table 19: Baseline data by IBD type.

	CD	UC/IBDU	Overall
	(N=1118)	(N=1026)	(N=2144)
Age			
Median (IQR)	41.5 (31.0 - 54.0)	48.0 (36.0 - 59.0)	44.0 (32.0 - 57.0)
Sex			
Male	478 (42.8%)	488 (47.6%)	966 (45.1%)
Female	640 (57.2%)	538 (52.4%)	1178 (54.9%)
Ethnicity			
White	849 (75.9%)	808 (78.8%)	1657 (77.3%)
Non-white	31 (2.8%)	26 (2.5%)	57 (2.7%)
Missing	238 (21.3%)	192 (18.7%)	430 (20.1%)
BMIcat			
Underweight	20 (1.8%)	18 (1.8%)	38 (1.8%)
Normal	451 (40.3%)	389 (37.9%)	840 (39.2%)
Overweight	379 (33.9%)	372 (36.3%)	751 (35.0%)
Obese	215 (19.2%)	216 (21.1%)	431 (20.1%)
Missing	53 (4.7%)	31 (3.0%)	84 (3.9%)
diagnosis			
CD	1118 (100%)	0 (0%)	1118 (52.1%)
UC	0 (0%)	950 (92.6%)	950 (44.3%)
IBDU	0 (0%)	76 (7.4%)	76 (3.5%)
IBD Duration			
Median (IQR)	10.3 (5.04 - 20.2)	9.28 (4.27 - 16.8)	9.80 (4.60 - 18.6)
Missing	64 (5.7%)	56 (5.5%)	120 (5.6%)
Smoke			
Current	63 (5.6%)	46 (4.5%)	109 (5.1%)
Previous	302 (27.0%)	325 (31.7%)	627 (29.2%)
Never	508 (45.4%)	457 (44.5%)	965 (45.0%)
Missing	245 (21.9%)	198 (19.3%)	443 (20.7%)
ECigs			
Current	39 (3.5%)	31 (3.0%)	70 (3.3%)
Previous	33 (3.0%)	20 (1.9%)	53 (2.5%)
Never	799 (71.5%)	776 (75.6%)	1575 (73.5%)
Missing	247 (22.1%)	199 (19.4%)	446 (20.8%)
as.numeric(IMD)			
Median (IQR)	4.00 (2.00 - 5.00)	4.00 (3.00 - 5.00)	4.00 (2.00 - 5.00)
Missing	11 (1.0%)	11 (1.1%)	22 (1.0%)
control_8			
Median (IQR)	13.0 (10.0 - 15.0)	13.0 (11.0 - 15.0)	13.0 (11.0 - 15.0)
Missing	248 (22.2%)	199 (19.4%)	447 (20.8%)
vas_control			
<85	318 (28.4%)	251 (24.5%)	569 (26.5%)
85+	547 (48.9%)	574 (55.9%)	1121 (52.3%)
Missing	253 (22.6%)	201 (19.6%)	454 (21.2%)
FC			
Median (IQR)	54.0 (20.0 - 167)	43.0 (20.0 - 155)	49.0 (20.0 - 161)
CReactiveProtein			
Median (IQR)	2.00 (1.00 - 5.00)	2.00 (1.00 - 5.00)	2.00 (1.00 - 5.00)

```
    PR02 +  
    Behaviour +  
    Perianal +  
    Surgery +  
    Smoke +  
    ECigs |  
    cat,  
    render.continuous = my.render.cont  
  )
```


	FC < 50 (N=533)	FC 50-250 (N=381)	FC > 250 (N=204)	Overall (N=1118)
IBD Duration				
Median (IQR)	10.2 (4.89 - 20.9)	10.6 (5.02 - 20.2)	10.1 (5.37 - 17.9)	10.3 (5.04 - 20.2)
Missing	33 (6.2%)	21 (5.5%)	10 (4.9%)	64 (5.7%)
Location				
L1	138 (25.9%)	99 (26.0%)	46 (22.5%)	283 (25.3%)
L2	128 (24.0%)	90 (23.6%)	60 (29.4%)	278 (24.9%)
L3	176 (33.0%)	136 (35.7%)	73 (35.8%)	385 (34.4%)
L4 only	3 (0.6%)	4 (1.0%)	3 (1.5%)	10 (0.9%)
Missing	88 (16.5%)	52 (13.6%)	22 (10.8%)	162 (14.5%)
L4				
Not present	398 (74.7%)	293 (76.9%)	151 (74.0%)	842 (75.3%)
Present	47 (8.8%)	36 (9.4%)	31 (15.2%)	114 (10.2%)
Missing	88 (16.5%)	52 (13.6%)	22 (10.8%)	162 (14.5%)
HBI				
Median (IQR)	2.00 (1.00 - 4.00)	2.00 (1.00 - 4.00)	2.00 (1.00 - 3.00)	2.00 (1.00 - 4.00)
Missing	331 (62.1%)	227 (59.6%)	116 (56.9%)	674 (60.3%)
PRO2				
Median (IQR)	0 (0 - 2.00)	0 (0 - 2.00)	0 (0 - 1.00)	0 (0 - 2.00)
Missing	242 (45.4%)	135 (35.4%)	69 (33.8%)	446 (39.9%)
Behaviour				
B1	279 (52.3%)	211 (55.4%)	127 (62.3%)	617 (55.2%)
B2	96 (18.0%)	73 (19.2%)	37 (18.1%)	206 (18.4%)
B3	51 (9.6%)	31 (8.1%)	11 (5.4%)	93 (8.3%)
Missing	107 (20.1%)	66 (17.3%)	29 (14.2%)	202 (18.1%)
Perianal				
No	283 (53.1%)	224 (58.8%)	112 (54.9%)	619 (55.4%)
Yes	149 (28.0%)	106 (27.8%)	66 (32.4%)	321 (28.7%)
Missing	101 (18.9%)	51 (13.4%)	26 (12.7%)	178 (15.9%)
Surgery				
No	262 (49.2%)	181 (47.5%)	118 (57.8%)	561 (50.2%)
Yes	251 (47.1%)	193 (50.7%)	78 (38.2%)	522 (46.7%)
Missing	20 (3.8%)	7 (1.8%)	8 (3.9%)	35 (3.1%)
Smoke				
Current	32 (6.0%)	26 (6.8%)	5 (2.5%)	63 (5.6%)
Previous	153 (28.7%)	108 (28.3%)	41 (20.1%)	302 (27.0%)
Never	234 (43.9%)	169 (44.4%)	105 (51.5%)	508 (45.4%)
Missing	114 (21.4%)	78 (20.5%)	53 (26.0%)	245 (21.9%)
ECigs				
Current	19 (3.6%)	18 (4.7%)	2 (1.0%)	39 (3.5%)
Previous	16 (3.0%)	13 (3.4%)	4 (2.0%)	33 (3.0%)
Never	382 (71.7%)	272 (71.4%)	145 (71.1%)	799 (71.5%)
Missing	116 (21.8%)	78 (20.5%)	53 (26.0%)	247 (22.1%)

Table of Crohn's disease variables by cohort

```
comp <- demo.cd
comp$cohort <- "All"

temp <- demo.cd %>%
  drop_na(cat)
temp$cohort <- "FC"

comp <- rbind(comp, temp)

temp <- subset(demo.cd, ParticipantNo %in% FFQ$participantno)
temp$cohort <- "FFQ"

comp <- rbind(comp, temp)

comp$cohort <- factor(comp$cohort,
  levels = c("All", "FC", "FFQ"),
  labels = c("Full cohort", "FC cohort", "FFQ cohort")
)

table1(
  ~ Location +
    L4 +
    HBI +
    PRO2 +
    Behaviour +
    Perianal +
    Surgery +
    Smoke +
    ECigs |
    cohort,
  data = comp,
  render.continuous = my.render.cont,
  overall = FALSE
)
```

	Full cohort	FC cohort	FFQ cohort
	(N=1370)	(N=1118)	(N=530)
Location			
L1	362 (26.4%)	283 (25.3%)	149 (28.1%)
L2	333 (24.3%)	278 (24.9%)	141 (26.6%)
L3	454 (33.1%)	385 (34.4%)	160 (30.2%)
L4 only	14 (1.0%)	10 (0.9%)	6 (1.1%)
Missing	207 (15.1%)	162 (14.5%)	74 (14.0%)
L4			
Not present	1031 (75.3%)	842 (75.3%)	412 (77.7%)
Present	132 (9.6%)	114 (10.2%)	44 (8.3%)
Missing	207 (15.1%)	162 (14.5%)	74 (14.0%)
HBI			
Median (IQR)	2.00 (1.00 - 4.00)	2.00 (1.00 - 4.00)	2.00 (1.00 - 3.00)
Missing	833 (60.8%)	674 (60.3%)	287 (54.2%)
PRO2			
Median (IQR)	0 (0 - 2.00)	0 (0 - 2.00)	1.00 (0 - 2.00)
Missing	555 (40.5%)	446 (39.9%)	179 (33.8%)
Behaviour			
B1	749 (54.7%)	617 (55.2%)	285 (53.8%)
B2	252 (18.4%)	206 (18.4%)	105 (19.8%)
B3	121 (8.8%)	93 (8.3%)	54 (10.2%)
Missing	248 (18.1%)	202 (18.1%)	86 (16.2%)
Perianal			
No	746 (54.5%)	619 (55.4%)	299 (56.4%)
Yes	397 (29.0%)	321 (28.7%)	151 (28.5%)
Missing	227 (16.6%)	178 (15.9%)	80 (15.1%)
Surgery			
No	684 (49.9%)	561 (50.2%)	249 (47.0%)
Yes	634 (46.3%)	522 (46.7%)	273 (51.5%)
Missing	52 (3.8%)	35 (3.1%)	8 (1.5%)
Smoke			
Current	73 (5.3%)	63 (5.6%)	32 (6.0%)
Previous	343 (25.0%)	302 (27.0%)	186 (35.1%)
Never	562 (41.0%)	508 (45.4%)	299 (56.4%)
Missing	392 (28.6%)	245 (21.9%)	13 (2.5%)
ECigs			
Current	45 (3.3%)	39 (3.5%)	17 (3.2%)
Previous	39 (2.8%)	33 (3.0%)	14 (2.6%)
Never	892 (65.1%)	799 (71.5%)	486 (91.7%)
Missing	394 (28.8%)	247 (22.1%)	13 (2.5%)

Associations between Crohn's Disease-only variables and cohort membership

None of the CD-only variables were found to significantly differ across cohorts.

Montreal location

```
pander(fisher.test(comp$Location, comp$cohort, workspace = 200000000))
```

Table 20: Fisher's Exact Test for Count Data: `comp$Location` and `comp$cohort`

P value	Alternative hypothesis
0.659	two.sided

Upper gastrointestinal inflammation

```
pander(chisq.test(comp$L4, comp$cohort))
```

Table 21: Pearson's Chi-squared test: `comp$L4` and `comp$cohort`

Test statistic	df	P value
1.616	2	0.4457

Harvey-Bradshaw index

```
pander(summary(aov(HBI ~ cohort, data = comp)))
```

Table 22: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	2.721	1.36	0.2382	0.788
Residuals	1221	6972	5.71	NA	NA

PRO2

```
pander(summary(aov(PRO2 ~ cohort, data = comp)))
```

Table 23: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	0.6806	0.3403	0.08516	0.9184
Residuals	1835	7333	3.996	NA	NA

Montreal behaviour

```
pander(chisq.test(comp$Behaviour, comp$cohort))
```

Table 24: Pearson's Chi-squared test: `comp$Behaviour` and `comp$cohort`

Test statistic	df	P value
1.81	4	0.7707

Perianal disease

```
pander(chisq.test(comp$Perianal, comp$cohort))
```

Table 25: Pearson's Chi-squared test: `comp$Perianal` and `comp$cohort`

Test statistic	df	P value
0.2153	2	0.898

Surgery

```
pander(chisq.test(comp$Surgery, comp$cohort))
```

Table 26: Pearson's Chi-squared test: `comp$Surgery` and `comp$cohort`

Test statistic	df	P value
2.961	2	0.2276

Smoking status

```
pander(chisq.test(comp$Smoke, comp$cohort))
```

Table 27: Pearson's Chi-squared test: `comp$Smoke` and `comp$cohort`

Test statistic	df	P value
1.023	4	0.9063

E-cigarette use

```
pander(chisq.test(comp$ECigs, comp$cohort))
```

Table 28: Pearson's Chi-squared test: `comp$ECigs` and `comp$cohort`

Test statistic	df	P value
3.416	4	0.4908

Table of ulcerative colitis/IBDU variables

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

demo.uc %>%
  drop_na(cat) %>%
  table1(
    x = ~ `IBD Duration` + Extent + Mayo + PRO2 + Smoke + ECigs | cat,
    render.continuous = my.render.cont
  )
```

	FC < 50 (N=544)	FC 50-250 (N=302)	FC > 250 (N=180)	Overall (N=1026)
IBD Duration				
Median (IQR)	10.1 (4.48 - 17.8)	8.30 (4.31 - 15.4)	7.45 (3.10 - 13.4)	9.28 (4.27 - 16.8)
Missing	27 (5.0%)	17 (5.6%)	12 (6.7%)	56 (5.5%)
Extent				
E1	74 (13.6%)	34 (11.3%)	20 (11.1%)	128 (12.5%)
E2	214 (39.3%)	126 (41.7%)	76 (42.2%)	416 (40.5%)
E3	126 (23.2%)	79 (26.2%)	48 (26.7%)	253 (24.7%)
Missing	130 (23.9%)	63 (20.9%)	36 (20.0%)	229 (22.3%)
Mayo				
Median (IQR)	0 (0 - 1.00)	0 (0 - 1.00)	0.500 (0 - 2.00)	0 (0 - 1.00)
Missing	171 (31.4%)	62 (20.5%)	58 (32.2%)	291 (28.4%)
PRO2				
Median (IQR)	0 (0 - 1.00)	0 (0 - 1.00)	0 (0 - 1.00)	0 (0 - 1.00)
Missing	172 (31.6%)	64 (21.2%)	59 (32.8%)	295 (28.8%)
Smoke				
Current	32 (5.9%)	10 (3.3%)	4 (2.2%)	46 (4.5%)
Previous	156 (28.7%)	105 (34.8%)	64 (35.6%)	325 (31.7%)
Never	253 (46.5%)	131 (43.4%)	73 (40.6%)	457 (44.5%)
Missing	103 (18.9%)	56 (18.5%)	39 (21.7%)	198 (19.3%)
ECigs				
Current	9 (1.7%)	15 (5.0%)	7 (3.9%)	31 (3.0%)
Previous	15 (2.8%)	3 (1.0%)	2 (1.1%)	20 (1.9%)
Never	417 (76.7%)	228 (75.5%)	131 (72.8%)	776 (75.6%)
Missing	103 (18.9%)	56 (18.5%)	40 (22.2%)	199 (19.4%)

Table of Ulcerative colitis/IBDU variables by cohort

```

comp <- demo.uc
comp$cohort <- "All"

temp <- demo.uc %>%
  drop_na(cat)
temp$cohort <- "FC"

comp <- rbind(comp, temp)

```

```

temp <- subset(demo.uc, ParticipantNo %in% FFQ$participantno)
temp$cohort <- "FFQ"

comp <- rbind(comp, temp)

comp$cohort <- factor(comp$cohort,
  levels = c("All", "FC", "FFQ"),
  labels = c("Full cohort", "FC cohort", "FFQ cohort")
)

table1(~ Extent + Mayo + PRO2 + Smoke + ECigs | cohort,
  data = comp,
  render.continuous = my.render.cont,
  overall = FALSE
)

```


	Full cohort (N=1259)	FC cohort (N=1026)	FFQ cohort (N=561)
Extent			
E1	166 (13.2%)	128 (12.5%)	81 (14.4%)
E2	503 (40.0%)	416 (40.5%)	250 (44.6%)
E3	318 (25.3%)	253 (24.7%)	130 (23.2%)
Missing	272 (21.6%)	229 (22.3%)	100 (17.8%)
Mayo			
Median (IQR)	0 (0 - 1.00)	0 (0 - 1.00)	0 (0 - 1.00)
Missing	366 (29.1%)	291 (28.4%)	136 (24.2%)
PRO2			
Median (IQR)	0 (0 - 1.00)	0 (0 - 1.00)	0 (0 - 1.00)
Missing	372 (29.5%)	295 (28.8%)	136 (24.2%)
Smoke			
Current	49 (3.9%)	46 (4.5%)	31 (5.5%)
Previous	378 (30.0%)	325 (31.7%)	213 (38.0%)
Never	509 (40.4%)	457 (44.5%)	298 (53.1%)
Missing	323 (25.7%)	198 (19.3%)	19 (3.4%)
ECigs			
Current	35 (2.8%)	31 (3.0%)	21 (3.7%)
Previous	22 (1.7%)	20 (1.9%)	13 (2.3%)
Never	878 (69.7%)	776 (75.6%)	508 (90.6%)
Missing	324 (25.7%)	199 (19.4%)	19 (3.4%)

Associations between UC/IBDU-only variables and cohort membership

None of the UC/IBDU-only variables were found to significantly differ across cohorts.

Montreal extent

```
pander(chisq.test(comp$Extent, comp$cohort))
```

Table 29: Pearson's Chi-squared test: `comp$Extent` and `comp$cohort`

Test statistic	df	P value
2.793	4	0.593

Mayo score

```
pander(summary(aov(Mayo ~ cohort, data = comp)))
```

Table 30: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	5.821	2.91	1.794	0.1665
Residuals	2050	3325	1.622	NA	NA

PRO2

```
pander(summary(aov(PRO2 ~ cohort, data = comp)))
```

Table 31: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cohort	2	1.319	0.6593	0.8346	0.4342
Residuals	2040	1611	0.7899	NA	NA

Smoking status

```
pander(chisq.test(comp$Smoke, comp$cohort))
```

Table 32: Pearson's Chi-squared test: comp\$Smoke and comp\$cohort

Test statistic	df	P value
0.398	4	0.9826

E-cigarette use

```
pander(chisq.test(comp$ECigs, comp$cohort))
```

Table 33: Pearson's Chi-squared test: `comp$ECigs` and `comp$cohort`

Test statistic	df	P value
0.02764	4	0.9999

```
demo %>%
  drop_na(cat) %>%
  saveRDS(paste0(outdir, "demo.RDS"))

demo %>%
  saveRDS(paste0(outdir, "demo-full.RDS"))

demo.cd %>%
  saveRDS(paste0(outdir, "demo-cd.RDS"))

demo.uc %>%
  saveRDS(paste0(outdir, "demo-uc.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): `utf8(v.1.2.4)`, `generics(v.0.1.3)`, `xml2(v.1.3.6)`, `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)`, `cellranger(v.1.1.0)`, `jsonlite(v.1.8.8)`, `Formula(v.1.2-5)`, `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)`, `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)`,

kableExtra(v.1.4.0), *curl(v.5.2.1)*, *vctrsv.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), *systemfonts(v.1.1.0)*, *xfun(v.0.44)*, *tidyselect(v.1.2.1)*, *rstudioapi(v.0.16.0)*,
htmltools(v.0.5.8.1), *svglite(v.2.1.3)*, *rmarkdown(v.2.27)* and *compiler(v.4.4.0)*

Licensed by CC BY unless otherwise stated.