

STA540_CS1

Raymond Sun

2026-01-28

Data Preprocessing

```
library(emmeans)
library(dplyr)
library(tidyverse)
CTN_dat = read_csv("CTN_FINAL.csv")
table(CTN_dat$Q6_1)
table(CTN_dat$Q6_3)
table(CTN_dat$Q6_4)
table(CTN_dat$Q6_5)
table(CTN_dat$Q6_6)
table(CTN_dat$Q6_7)
table(CTN_dat$PO_FLAG)
CTN_filt = CTN_dat |>
  filter(PO_FLAG == "Include")
table(CTN_filt$FINISHED)
CTN_final = CTN_filt |>
  filter(SITE != "Yahoo")
```

255 of the 271 observations have their primary outcome analysis flag listed as “Include”, and the rest are excluded from the final dataset. Among these 255, 1 observation has the site listed as “Yahoo” which does not fit the criteria for this study. With this observation excluded, the remaining 254 observations are consistent with the manuscript and included in the final dataset `CTN_Final`.

Table 1

Age in years, median (IQR):

```
median(CTN_final$Q3_1)
```

[1] 25

```
quantile(CTN_final$Q3_1, 0.25)
```

25%
21

```
quantile(CTN_final$Q3_1, 0.75)
```

75%
27

The manuscript has the IQR as 23-27, so the 25th percentile isn't consistent with what I have reproduced (21).

Ethnicity, n (%):

```
CTN_final$hispanic <- ifelse(CTN_final$Q5_1 == 1,  
                           "Hispanic/Latino",  
                           "Not Hispanic/Latino")  
tab_eth <- table(CTN_final$hispanic)  
percent <- prop.table(tab_eth) * 100  
data.frame(  
  n = tab_eth["Hispanic/Latino"],  
  percent = round(percent["Hispanic/Latino"], 1)  
)
```

	n	percent
Hispanic/Latino	66	26

Race, n (%):

```
race_recoded <- ifelse(  
  CTN_final$Q5_3 == 23, "White",  
  ifelse(  
    CTN_final$Q5_3 == 24, "Black or African American",  
    ifelse(  
      ...
```

```

    CTN_final$Q5_3 == 25, "American Indian or Alaskan Native",
  ifelse(
    CTN_final$Q5_3 == 28, "Other",
    "Multiracial"
  )
)
)
)

tab_race_merged <- table(race_recoded)

cbind(
  n = as.vector(tab_race_merged),
  percent = round(100 * prop.table(tab_race_merged), 1)
)

```

	n	percent
American Indian or Alaskan Native	1	0.4
Black or African American	196	78.4
Multiracial	11	4.4
Other	14	5.6
White	28	11.2

History of PrEPa uptake, n (%):

```

prep_recoded <- ifelse(
  CTN_final$Q6_2 == 3, "Never taken PrEP",
  ifelse(
    CTN_final$Q6_2 == 1, "In the past 6 months",
    NA
  )
)

tab_prep <- table(prep_recoded, useNA = "no")

cbind(
  n = tab_prep,
  percent = round(prop.table(tab_prep) * 100, 1)
)

```

n	percent
---	---------

In the past 6 months	22	8.7
Never taken PrEP	232	91.3

The manuscript seems to have mistakenly listed the percentage for “In the past 6 months” as 8.9% here, but it should be 8.7% as that wouldn’t add up to 100%.

Number of male sex partners in the past 90 days, median (IQR):

```
median(CTN_final$Q11_2)
```

[1] 4

```
quantile(CTN_final$Q11_2, 0.25)
```

25%
3

```
quantile(CTN_final$Q11_2, 0.75)
```

75%
6

Condom use, n (%):

```
q11_recoded <- factor(
  CTN_final$Q11_3,
  levels = c(1, 2, 3, 4, 5),
  labels = c(
    "Never",
    "Sometimes",
    "About half the time",
    "Most of the time",
    "Always"
  )
)

tab_q11 <- table(q11_recoded, useNA = "no")

cbind(
  n = tab_q11,
  percent = round(prop.table(tab_q11) * 100, 1)
)
```

	n	percent
Never	36	14.2
Sometimes	108	42.5
About half the time	37	14.6
Most of the time	68	26.8
Always	5	2.0

The percentage for “About half the time” is 14.5 on the manuscript but rounded to 14.6 here.

Condomless receptive anal sex in the past 90 days, n (%):

```
CTN_final$condomless_RAS <- ifelse(
  CTN_final$Q11_4 == 1,
  "Condomless receptive anal sex in past 90 days",
  "No"
)

tab_ras <- table(CTN_final$condomless_RAS)
percent <- prop.table(tab_ras) * 100

data.frame(
  n = tab_ras["Condomless receptive anal sex in past 90 days"],
  percent = round(percent["Condomless receptive anal sex in past 90 days"], 1)
)
```

	n	percent
Condomless receptive anal sex in past 90 days	210	82.7

Ever tested for HIV during lifetime, n (%)

```
# Recode Q11_5
CTN_final$ever_HIV_test <- ifelse(
  CTN_final$Q11_5 == 1,
  "Ever tested for HIV during lifetime",
  "No"
)

# Create table
tab_HIV <- table(CTN_final$ever_HIV_test)
percent <- prop.table(tab_HIV) * 100

# Display only the 'Yes' row
```

```
data.frame(
  n = tab_HIV["Ever tested for HIV during lifetime"],
  percent = round(percent["Ever tested for HIV during lifetime"], 1)
)
```

```
          n percent
Ever tested for HIV during lifetime 191    75.2
```

```
CTN_tested = CTN_final |>
  filter(Q11_5 == 1)
round(median(CTN_tested$LAST_HIV_TEST_MONTHS, na.rm = TRUE))
```

```
[1] 11
```

```
round(quantile(CTN_tested$LAST_HIV_TEST_MONTHS, 0.25, na.rm = TRUE))
```

```
25%
6
```

```
round(quantile(CTN_tested$LAST_HIV_TEST_MONTHS, 0.75, na.rm = TRUE))
```

```
75%
21
```

If not tested for HIV, n (%):

```
data.frame(
  n = tab_HIV["No"],
  percent = round(percent["No"], 1)
)
```

```
          n percent
No 63    24.8
```

Main reasons cited by the 63 participants for not getting tested, n (%):

```

CTN_no = CTN_final |>
  filter(Q11_5 == 2)
q11_7_recoded <- factor(
  CTN_no$Q11_7,
  levels = c(1, 2, 3, 4, 5, 6, 7, 8),
  labels = c(
    "Unlikely to be exposed to HIV",
    "Afraid of testing HIV-positive",
    "Did not want to think about HIV/HIV-positive",
    "Worried about names being reported if positive",
    "Dislike for needles",
    "Unable to trust that the results will be confidential",
    "Unaware of where to get tested",
    "Other reason"
  )
)

tab_q11_7 <- table(q11_7_recoded, useNA = "no")
cbind(
  n = tab_q11_7,
  percent = round(prop.table(tab_q11_7) * 100, 1)
)

```

	n	percent
Unlikely to be exposed to HIV	8	12.7
Afraid of testing HIV-positive	26	41.3
Did not want to think about HIV/HIV-positive	8	12.7
Worried about names being reported if positive	3	4.8
Dislike for needles	5	7.9
Unable to trust that the results will be confidential	3	4.8
Unaware of where to get tested	7	11.1
Other reason	3	4.8

Primary Analysis

EDA:

```

# CTN_ord is a dataframe of all participants who ordered a self-test kit
CTN_ord = CTN_final |>
  filter(ORA_REDEEMED == "Yes")
table(CTN_ord$PLATFORM)

```

```
Dating app      Info site Social media
    134           17          26
```

```
table(CTN_ord$SITE_TYPE)
```

```
Dating Apps   Info Sites Social Media
    138           17          22
```

```
table(CTN_ord$WAVE)
```

```
1   2   4
9 138 30
```

```
CTN_rec <- CTN_ord
CTN_rec$WAVE <- ifelse(
  CTN_rec$WAVE %in% c(1, 4),
  1, # Manuscript Wave 1 (original + second phase)
  2 # Manuscript Wave 2
)
```

Since the total number of kits ordered in wave 1 & 4 (39) corresponds to the total number of kits in wave 1 and the second phase of wave 1 in the manuscript, it is reasonable to assume wave 4 is the second phase of wave 1, so it is recoded as wave 1. There are 4 participants recruited from Instagram with platform listed as “Social media” but site_type listed as “Dating Apps”, so it makes sense to use the platform variable which aligns with table 2 of the manuscript.

```
CTN_counts <- CTN_rec %>%
  count(WAVE, PLATFORM, name = "kits") |>
  complete(
    WAVE = c(1, 2), # numeric for complete()
    PLATFORM = c("Dating app", "Social media", "Info site"),
    fill = list(kits = 0)
  ) |>
  mutate(
    time = ifelse(WAVE == 1, 70, 38), # numeric comparison
    kits_per_day = round(kits / time, 2),
    kits_adj = kits + 0.5,
```

```

WAVE = factor(WAVE, levels = c(1,2)),
PLATFORM = factor(PLATFORM, levels = c("Dating app", "Social media", "Info site"))
)
pois_mod <- glm(
  kits ~ WAVE * PLATFORM,
  family = poisson,
  offset = log(time),
  data = CTN_counts
)
emm = emmeans(
  pois_mod,
  ~ PLATFORM | WAVE,
  type = "response",
  offset = 0
)

```

NOTE: A nesting structure was detected in the fitted model:
 WAVE %in% .static.offset.

```

emm_round <- as.data.frame(emm) %>%
  mutate(across(where(is.numeric), ~ round(.x, 2)))
emm_round

```

	WAVE .static.offset.	PLATFORM	rate	SE	df	asymp.LCL	asymp.UCL	
2	2	3.64	Dating app	3.29	0.29	Inf	2.76	3.92
3	1	4.25	Dating app	0.13	0.04	Inf	0.07	0.25
6	2	3.64	Social media	0.34	0.09	Inf	0.20	0.59
7	1	4.25	Social media	0.19	0.05	Inf	0.11	0.32
10	2	3.64	Info site	0.00	0.00	Inf	0.00	Inf
11	1	4.25	Info site	0.24	0.06	Inf	0.15	0.39

The rate estimations from the Poisson regression model successfully reproduced the rates in the manuscript: Jack'D - 3.29 kits/day, Instagram - 0.34 kits/day, Bing - 0 kits/day, Grindr - 0.13 kits/day, Facebook - 0.19 kits/day, Google - 0.24 kits/day.

```

pw <- contrast(
  emm,
  method = "pairwise", # differences
  adjust = "none"        # adjust later, per wave
)

```

```

pw_df <- as.data.frame(pw)
wave1 <- pw_df %>%
  filter(WAVE == 1) %>%
  mutate(p_adj = p.adjust(p.value, method = "BH")) |>
  mutate(across(where(is.numeric), ~ round(.x, 2)))

```

```

# Pairwise contrasts with BH adjustment and rounding
wave2 <- pw_df %>%
  filter(WAVE == 2) %>%
  mutate(p_adj = p.adjust(p.value, method = "BH")) |>
  mutate(across(where(is.numeric), ~ round(.x, 2)))

```

```
wave1
```

	.static.offset.	WAVE	contrast	ratio	SE	df	null	z.ratio
1	4.24849524204936	1	Dating app / Social media	0.69	0.30	Inf	1	-0.85
2	4.24849524204936	1	Dating app / Info site	0.53	0.22	Inf	1	-1.54
3	4.24849524204936	1	Social media / Info site	0.76	0.28	Inf	1	-0.73
		p.value	p_adj					
1		0.40	0.47					
2		0.12	0.37					
3		0.47	0.47					

```
wave2
```

	.static.offset.	WAVE	contrast	ratio	SE	df
1	3.63758615972639	2	Dating app / Social media	9.620000e+00	2.800000e+00	Inf
2	3.63758615972639	2	Dating app / Info site	6.064565e+11	2.562107e+16	Inf
3	3.63758615972639	2	Social media / Info site	6.307148e+10	2.664591e+15	Inf
		null	z.ratio	p.value	p_adj	
1	1	7.77	0	0		
2	1	0.00	1	1		
3	1	0.00	1	1		

```

pois_adj = glm(
  kits_adj ~ WAVE * PLATFORM,
  family = poisson,
  offset = log(time),
  data = CTN_counts
)
emm_adj = emmeans(

```

```

  pois_adj,
  ~ PLATFORM | WAVE,
  type = "response",
  offset = 0
)

```

NOTE: A nesting structure was detected in the fitted model:
 WAVE %in% .static.offset.

```

pw_adj = contrast(
  emm_adj,
  method = "pairwise", # differences
  adjust = "none"       # adjust later, per wave
)
pw_df_adj <- as.data.frame(pw_adj)
wave2_adj <- pw_df_adj |>
  filter(WAVE == 2) |>
  mutate(p_adj = p.adjust(p.value, method = "BH")) %>%
  mutate(across(where(is.numeric), ~ round(.x, 2)))
wave2_adj

```

	.static.offset. WAVE	contrast	ratio	SE	df	null	z.ratio
1	3.63758615972639	2 Dating app / Social media	9.3	2.66	Inf	1	7.78
2	3.63758615972639	2 Dating app / Info site	251.0	355.67	Inf	1	3.90
3	3.63758615972639	2 Social media / Info site	27.0	38.88	Inf	1	2.29
	p.value	p_adj					
1	0.00	0.00					
2	0.00	0.00					
3	0.02	0.02					

For the contrasts, the adjusted p-values for wave 1 (0.47, 0.37, 0.47) are slightly lower than in the manuscript (0.59), while results for wave 2 could be reproduced by artificially adding 0.5 to the amount of kits ordered so that zero rates are avoided, since a zero count would cause problems for the Poisson regression model. However, there isn't any information that suggests this in the manuscript or appendix, so I believe that the p-values are generally not reproducible. This could be attributed to differences between `emmeans` in R and SAS code, or some technical procedures that were used on the data but not reported in the manuscript.

Secondary Analysis

Fisher's exact test for association between "People in my life would leave if I had HIV":

```
CTN_final <- CTN_final |>
  mutate(
    ORDERED = ifelse(ORA_REDEEMED == "Yes", 1, 0)
  )
tab <- table(CTN_final$Q15_5, CTN_final$ORDERED)

dimnames(tab) <- list(
  Ordered = c("No", "Yes"),
  Statement = c("Agree", "Disagree")
)

fisher.test(tab)
```

```
Fisher's Exact Test for Count Data

data: tab
p-value = 0.03517
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
1.013692 3.252028
sample estimates:
odds ratio
1.814215
```

Wilcoxon test for "I think that new HIV/AIDS treatments can eradicate the virus from your body" and "I could not be friends with someone who has HIV/AIDS":

```
wilcox.test(
  Q94_13 ~ ORDERED,
  data = CTN_final,
  exact = FALSE
)
```

```
Wilcoxon rank sum test with continuity correction
```

```
data: Q94_13 by ORDERED  
W = 4204.5, p-value = 0.02897  
alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(  
  Q14_3 ~ ORDERED,  
  data = CTN_final,  
  exact = FALSE  
)
```

```
Wilcoxon rank sum test with continuity correction  
  
data: Q14_3 by ORDERED  
W = 5878, p-value = 0.03248  
alternative hypothesis: true location shift is not equal to 0
```

The Fisher exact test successfully reproduces the secondary result given in the original study (0.035), while the Wilcoxon tests also nearly replicate the results in the original study (0.029 and 0.033), the difference likely comes from rounding.