Case Study #3: Efficacy of Combination Intervention in Botswana to combat HIV infection

Anisha Jain

2024-12-10

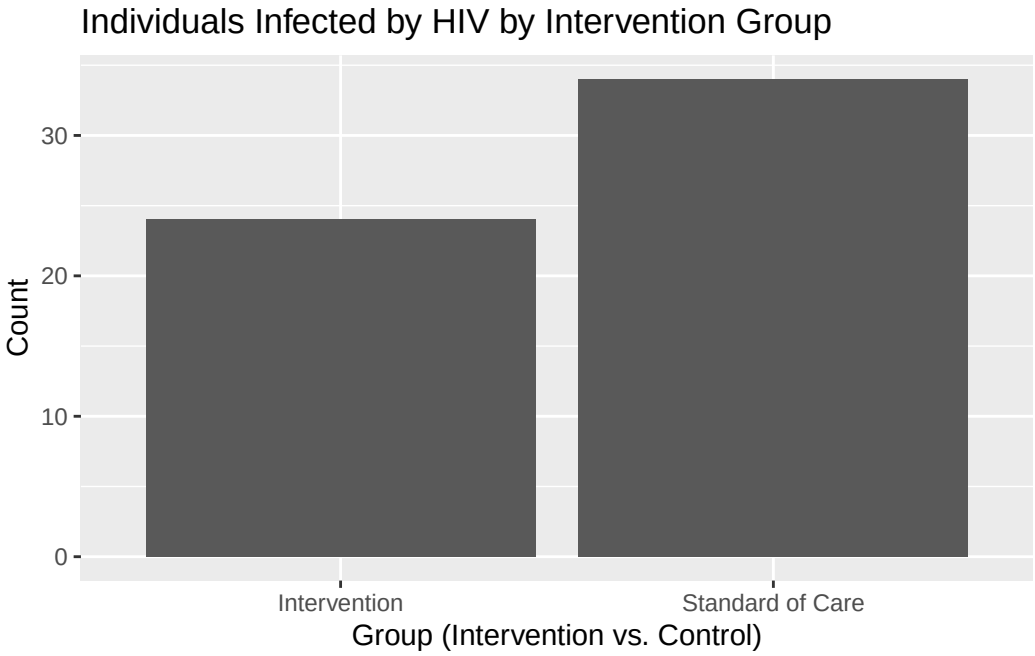
Introduction

The Botswana Combination Prevention Project (BCPP) conducted a clinical trial that applied a community-based approach against HIV in Botswana. Fifteen communities out of 30 were randomly selected to have this intervention take place, and after one year BCPP followed up on 7,826 individuals HIV infection status. These individuals were originally HIV-negative. In this case study, we are investigating the effect of the intervention on these 7,826 individuals, both before and after adjusting for important baseline information. Here, our response variable is the HIV status after one year, and the explanatory variable is the intervention strategy. Through this report, we hope to determine if the intervention was effective for reducing the probability of contracting HIV to inform future intervention strategies.

Exploratory Data Analysis

This data set comes from the Botswana Combination Prevention Project (BCPP) and contains 7826 observations. Two variables, whether alcohol and whether a condom was used in the participants last sexual encounter, were removed due to a high number (approximately 12.5%) of participants without a response. Additionally, a small number (approximately 3.5%) of all participants were missing data on their level of community engagement. After removing these participants, the dataset contained 7551 observations.

Type	n	number_infected	number_uninfected
Intervention	3714	24	3690
Standard of Care	3837	34	3803



From the summary figures, we see that the large majority of participants remained uninfected with HIV after the one-year period. We also notice that more individuals in the baseline standard of care group were infected, but it is unclear whether the difference in the two groups is statistically significant from the figures alone.

Methods

We used a logistic regression model to predict the log-odds of infection $logit(\pi)$ from the the intervention treatment $intervention$. We used the baseline variables gender ($male$), marital status ($married, widowed, single$), and employment status ($looking, notlooking$), to adjust our model. Our model for the population without adjusting for these variables is

$$logit(\pi(intervention_i)) = \beta_0 + \beta_1(intervention_i).$$

Our model for the population when adjusting for these variables is

$$logit(\pi(intervention_i, male_i, married_i, widowed_i, single_i, looking_i, notlooking_i)) = \beta_0 + \beta_1(intervention_i)$$

$$+\beta_2(male_i) + \beta_3(married_i) + \beta_4(widowed_i) + \beta_5(single_i) + \beta_6(looking_i) + \beta_7(notlooking_i).$$

The variables chosen for adjustment were chosen through forward automated variable selection using AIC as the selection metric. The variables that we are adjusting by had low AIC through the automated selection process, meaning that adding them allowed us to explain enough variablity in the data to compensate for increasing the complexity of the model. More sociologically, gender, marital status, and employment status can provide a general picture of lifestyle and socioeconomic status, and add meaningfully to the model.

Results

The estimated model parameters, their standard errors, 95% confidence intervals, and p-value can be found below for the adjusted model and the simple model. Note that the confidence intervals and parameters are reported on the log-odds scale.

Table 2: Coefficients and Standard Errors for Adjusted Model

	Estimate	Std_Error	Lower	Upper	P_Value
(Intercept)	-4.226	0.777	-5.749	-2.703	0.000
random_armStandard of Care	0.256	0.269	-0.271	0.783	0.341
genderM	-0.939	0.326	-1.578	-0.299	0.004
marital_statusMarried	-2.108	0.920	-3.911	-0.304	0.022
marital_statusSingle/never married	-0.798	0.732	-2.232	0.636	0.275
marital_statusWidowed	-14.798	451.537	-899.795	870.199	0.974
employment_statusUnemployed looking for work	0.729	0.363	0.016	1.441	0.045
employment_statusUnemployed not looking for work	0.300	0.416	-0.516	1.115	0.471

Table 3: Coefficients and Standard Errors for the simple model

	Estimate	Std_Error	Lower	Upper	P_Value
(Intercept)	-5.035	0.205	-5.437	-4.634	0.000
random_armStandard of Care	0.318	0.268	-0.206	0.843	0.234

In the simple model, we are 95% confident that using the standard of care instead of the intervention is associated with between a .814 and 3.323 times change in the odds of contracting HIV. In the adjusted model, we are 95% confident that, holding all other variables constant, using the standard of care instead of the intervention is associated with between a .763 and 2.188 times change in the odds of contracting HIV.

In the simple model, we found no statistically significant evidence that the intervention associates with the odds of becoming HIV-infected ($\alpha = .05$). Additionally, in the adjusted model, given the gender, marital status, and employment status of the participant, the intervention variable is not statistically significant in the model.

Conclusion

This report investigated the efficacy of an intervention done by the BCPP. We found that there is no statistically significant evidence that the intervention associates with the odds of becoming HIV-infected, whether or not we adjust for the gender, marital status, and employment status. This indicates that the organizers of BCPP should consider another approach to combat HIV-infection in Botswana. Since the data is coming from only 30 specific communities it could be interesting to analyze participants at the community level. It could be that certain communities had better

care practices in place to begin with, or others are more receptive to the interventions brought by BCPP. Since the 30 communties could be considered clusters, analyzing at the community level could remove any inter-community dependence introduced.

R Appendix

```
# Loading necessary packages
library(tidyverse)
library(Sleuth2)      # the package containing the data for the case study
library(kableExtra)   # for creating nicely formatted tables in Quarto
library(performance)
library(see)
library(broom)

# Loading the case study data
bcpp <- read_csv("bcpp-year-one.csv")
```

```
Rows: 7826 Columns: 17
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (15): de_subj_idC, random_arm, hiv_status_current, gender, community_eng...
dbl  (2): age_at_interview, partners_lifetime

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
bcpp <- bcpp |> mutate(hiv_stats = hiv_status_current =="HIV-infected")

#remove participants with no community engagement var
bcpp_clean <- bcpp |> filter(!is.na(community_engagement))
bcpp_clean$condom_lastsex <- NULL
bcpp_clean$alcohol_lastsex <- NULL

#showing which data are missing
apply(bcpp_clean, 2, FUN=function(x)mean(is.na(x)) * 100)
```

```
         de_subj_idC           random_arm   hiv_status_current
                   0                    0                    0
              gender     age_at_interview community_engagement
                   0                    0                    0
     religious_affil       marital_status            education
                   0                    0                    0
   employment_status    partners_lifetime           dx_tb_ever
```

```
                  0                   0                   0
dx_heartdisease_ever             alcohol               smoke
                  0                   0                   0
           hiv_stats
                  0
```

```
apply(bcpp, 2, FUN=function(x)mean(is.na(x)) * 100)
```

```
         de_subj_idC           random_arm   hiv_status_current
            0.000000             0.000000             0.000000
              gender     age_at_interview community_engagement
            0.000000             0.000000             3.513928
     religious_affil       marital_status            education
            0.000000             0.000000             0.000000
   employment_status    partners_lifetime       condom_lastsex
            0.000000             0.000000            12.484028
     alcohol_lastsex           dx_tb_ever dx_heartdisease_ever
           12.407360             0.000000             0.000000
             alcohol                 smoke            hiv_stats
            0.000000             0.000000             0.000000
```

```
#conducting the forward automated variable selection
full_glm <- glm(hiv_stats~random_arm +
 ↪  gender+age_at_interview+community_engagement+religious_affil+marital_status+education+emplo
 ↪  +dx_tb_ever+dx_heartdisease_ever, family='binomial', data=bcpp_clean)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
simple_glm <- glm(hiv_stats~random_arm, family='binomial', data=bcpp_clean)

step(simple_glm, scope = list(lower = simple_glm, upper = full_glm),
 ↪  direction="forward")
```

```
Start:  AIC=682.92
hiv_stats ~ random_arm

                      Df Deviance    AIC
+ gender               1   669.84 675.84
+ marital_status       3   666.84 676.84
+ employment_status    2   671.53 679.53
+ education            4   667.72 679.72
+ age_at_interview     1   676.32 682.32
<none>                     678.92 682.92
+ dx_tb_ever           1   677.34 683.34
```

```
+ dx_heartdisease_ever  1   677.80 683.80
+ partners_lifetime     1   678.47 684.47
+ religious_affil       1   678.80 684.80
+ community_engagement  2   677.17 685.17

Step:  AIC=675.84
hiv_stats ~ random_arm + gender

                       Df Deviance    AIC
+ marital_status        3   656.23 668.23
+ education             4   658.20 672.20
+ employment_status     2   663.40 673.40
+ age_at_interview      1   666.17 674.17
<none>                      669.84 675.84
+ dx_tb_ever            1   668.30 676.30
+ religious_affil       1   668.31 676.31
+ dx_heartdisease_ever  1   668.59 676.59
+ partners_lifetime     1   669.82 677.82
+ community_engagement  2   667.96 677.96

Step:  AIC=668.23
hiv_stats ~ random_arm + gender + marital_status

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

                       Df Deviance    AIC
+ employment_status     2   651.26 667.26
+ education             4   647.42 667.42
<none>                      656.23 668.23
+ dx_tb_ever            1   654.82 668.82
+ dx_heartdisease_ever  1   655.23 669.23
+ religious_affil       1   655.47 669.47
+ age_at_interview      1   655.95 669.95
+ partners_lifetime     1   656.21 670.21
+ community_engagement  2   654.65 670.65

Step:  AIC=667.26
hiv_stats ~ random_arm + gender + marital_status + employment_status

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

                       Df Deviance    AIC
<none>                      651.26 667.26
+ dx_tb_ever            1   649.92 667.92
```

```
+ education             4   644.08 668.08
+ dx_heartdisease_ever  1   650.33 668.33
+ religious_affil       1   650.79 668.79
+ age_at_interview      1   651.06 669.06
+ partners_lifetime     1   651.26 669.26
+ community_engagement  2   649.55 669.55
```

```
Call:  glm(formula = hiv_stats ~ random_arm + gender + marital_status +
    employment_status, family = "binomial", data = bcpp_clean)

Coefficients:
                                        (Intercept)
                                            -4.2257
                          random_armStandard of Care
                                             0.2559
                                            genderM
                                            -0.9389
                               marital_statusMarried
                                            -2.1076
                  marital_statusSingle/never married
                                            -0.7978
                               marital_statusWidowed
                                           -14.7982
    employment_statusUnemployed looking for work
                                             0.7285
employment_statusUnemployed not looking for work
                                             0.2998

Degrees of Freedom: 7550 Total (i.e. Null);  7543 Residual
Null Deviance:      680.4
Residual Deviance: 651.3    AIC: 667.3
```

```r
#we found that hiv_stats ~ random_arm + gender + marital_status +
 ↪  employment_status has lowest AIC
adjusted_glm <- glm(hiv_stats~random_arm + gender + marital_status +
 ↪  employment_status, family='binomial', data=bcpp_clean)
```

```r
# creating a table for the descriptive statistics for hiv status based on
 ↪  intervention
status_intervention <- bcpp_clean |> filter(random_arm == "Intervention") |>
  summarize(Type = "Intervention",
            n = n(),
            number_infected = sum(hiv_status_current == "HIV-infected", na.rm =
             ↪  TRUE),
            number_uninfected = sum(hiv_status_current == "HIV-uninfected", na.rm
             ↪  = TRUE))
```

```
status_soc <- bcpp_clean |> filter(random_arm == "Standard of Care") |>
  summarize(Type = "Standard of Care",
            n = n(),
            number_infected = sum(hiv_status_current == "HIV-infected", na.rm =
             ↪  TRUE),
            number_uninfected = sum(hiv_status_current == "HIV-uninfected", na.rm
             ↪  = TRUE))


# joining the tables of descriptive statistics
status_table <- status_intervention |>
  full_join(status_soc, by = c("Type", "n", "number_infected",
   ↪  "number_uninfected"))
```

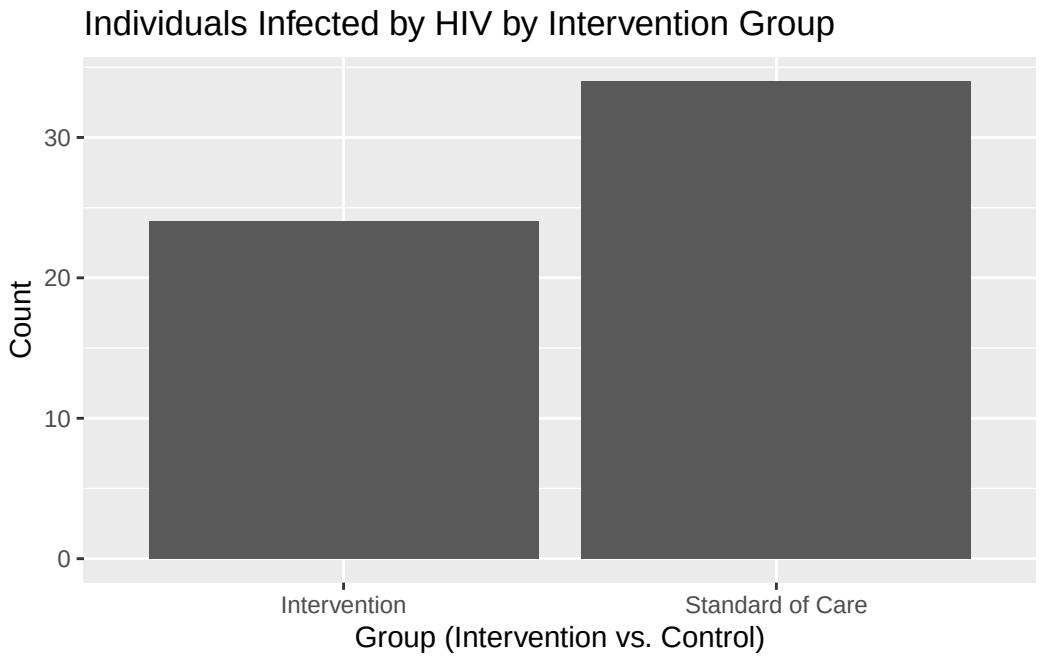
```
kable(status_table)
```

Type	n	number_infected	number_uninfected
Intervention	3714	24	3690
Standard of Care	3837	34	3803

```
#creating visual of the info just for those who are infected
hiv_summary <- bcpp_clean |>
  group_by(random_arm, hiv_status_current) |>
  summarize(count = n(), .groups = "drop")

# filtering for HIV infected
hiv_infected <- hiv_summary |> filter(hiv_status_current == "HIV-infected")

# Create the bar chart
ggplot(hiv_infected, aes(x = random_arm, y = count)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Individuals Infected by HIV by Intervention Group",
    x = "Group (Intervention vs. Control)",
    y = "Count"
  )
```




```
# Adjusted glm - Representing the regression table as a dataframe (i.e., tidying
 ↪  the summary() output)
z = qnorm(1- .05/2)
model_summary <- summary(adjusted_glm)
coefficients_table <- data.frame(
  Estimate = round(coef(model_summary)[, "Estimate"], 3),
  Std_Error = round(coef(model_summary)[, "Std. Error"], 3),
  Lower = round(coef(model_summary)[, "Estimate"] - z * coef(model_summary)[,
   ↪  "Std. Error"], 3),
  Upper = round(coef(model_summary)[, "Estimate"] + z * coef(model_summary)[,
   ↪  "Std. Error"], 3),
  P_Value = round(coef(model_summary)[, "Pr(>|z|)"], 3)
)
```

Table 5: Coefficients and Standard Errors for Adjusted Model

	Estimate	Std_Error	Lower	Upper	P_Value
(Intercept)	-4.226	0.777	-5.749	-2.703	0.000
random_armStandard of Care	0.256	0.269	-0.271	0.783	0.341
genderM	-0.939	0.326	-1.578	-0.299	0.004
marital_statusMarried	-2.108	0.920	-3.911	-0.304	0.022
marital_statusSingle/never married	-0.798	0.732	-2.232	0.636	0.275
marital_statusWidowed	-14.798	451.537	-899.795	870.199	0.974
employment_statusUnemployed looking for work	0.729	0.363	0.016	1.441	0.045
employment_statusUnemployed not looking for work	0.300	0.416	-0.516	1.115	0.471

```r
# Simple glm - Representing the regression table as a dataframe (i.e., tidying
 ↪  the summary() output)

z = qnorm(1- .05/2)
model_summary <- summary(simple_glm)
coefficients_table_simple <- data.frame(
  Estimate = round(coef(model_summary)[, "Estimate"], 3),
  Std_Error = round(coef(model_summary)[, "Std. Error"], 3),
  Lower = round(coef(model_summary)[, "Estimate"] - z * coef(model_summary)[,
   ↪  "Std. Error"], 3),
  Upper = round(coef(model_summary)[, "Estimate"] + z * coef(model_summary)[,
   ↪  "Std. Error"], 3),
  P_Value = round(coef(model_summary)[, "Pr(>|z|)"], 3)
)
```

Table 6: Coefficients and Standard Errors for the simple model

	Estimate	Std_Error	Lower	Upper	P_Value
(Intercept)	-5.035	0.205	-5.437	-4.634	0.000
random_armStandard of Care	0.318	0.268	-0.206	0.843	0.234