BCPP Preliminary Analysis

Melody Owen

2025-05-30

OVERVIEW

The Botswana Combination Prevention Project (BCPP)

Motivation

- Goal: The primary goal of BCPP was to determine whether implementation of combination prevention package (CP) can significantly reduce population-level, cumulative HIV incidence
- Population: Individuals in Botswana aged 16-64 years
- Timeline: Study length was approximately 3 years
- Design: 30 communities were selected and matched into pairs based on community characteristics thought to be associated with HIV incidence

Treatment in BCPP

The Combination Prevention (CP) prevention package included the following four components:

1. VMMC: Male circumcision (only for HIV-negative males)
2. HTC: HIV Testing and Counseling (only for HIV-negative individuals)
3. ART: Antiretroviral Therapy (only for HIV-positive individuals)
4. PMTCT: Prevention of mother-to-child transmission (only for pregnant HIV-positive females)

Clusters (30 communities) were randomized to either:

- Treatment: CP Package
- Control: Standard of Care

Our analysis examines the impact of CP for HIV-negative individuals, so we consider the "entire" package components 1 and 2 only.

Questions of Interest

1. What is the direct effect (OR) of the CP intervention?
    a. This is the direct effect among those to intervention villages
    b. Will be a standard mediation analysis
    c. Related: To what extent is the effect mediated by receipt of VMMC? Look at the mediation proportion in a standard mediation analysis.
2. To what extent is the total effect mediated by VMMC (Voluntary Male Medical Circumcision)?
    a. In this case, "overall" means combining spillover effects with individual effects
    b. Define the mediation proportion
3. To what extent is the individual effect of CP mediated by VMMC?
4. What is the overall effect (OR) of CP on HIV incidence?
    a. Use a logistic regression model
    b. Intent to treat (ITT) analysis
5. What is the spillover effect (OR) of CP in intervention villages compared to those who did not take up CP?
    a. Spillover effect is among those in intervention villages who did not receive the intervention
    b. Related: To what extent is the spillover effect mediated by villages level VMMC deliver (mediation proportion)? (Look at Tyler's paper)
6. What is the total spillover (OR) of the CP intervention?
    a. Total spillover effect is everyone in BCPP who did not receive the intervention in control and intervention villages
    b. Related: To what extent is the total spillover effect mediated by village level of VMMC delivery (mediation proportion)?
7. What is the total effect of CP across all villages in the study?

NOTATION

K is the total number of villages in the study, indexed as $k = 1, ..., K$

m_k is the total number of individuals in cluster k, indexed as $i = 1, ..., m_k$

- $m_k^{(\text{male})}$ are the total number of males in cluster k
- $m_k^{(\text{female})}$ are the total number of females in cluster k

Y_{ik} is the outcome of subject i in cluster k, and is binary

- In BCPP, $Y_{ik} = 1$ if a subject seroconverted by the end of the study, $Y_{ik} = 0$ otherwise

T_k is the cluster-level binary treatment assignment

- In BCPP, $T_k = 1$ if a cluster has been assigned to receive CP, and $T_k = 0$ otherwise

$X_{ik}^{(1)}$, $X_{ik}^{(2)}$ denotes each of the two components of the treatment, T_k.

- In BCPP, the Combination Prevention (CP) package included the following:
  1. MC: Male Circumcision (available only for HIV-negative males)
  2. HTC: HIV Testing and Counseling (available only for HIV-negative individuals)
  3. ART: Antiretroviral Therapy (available only for HIV-positive individuals)
  4. PMTCT: Prevention of Mother-to-Child Transmission (available only for HIV-positive females)
- We are only considering the first two components as the entire treatment package, since the last two apply to HIV-positive individuals only.
- $X_{ik}^{(1)}$ = "Yes" if individual i in cluster k was circumcised before or during the study, $X_{ik}^{(1)}$ = "No" if they are male and not circumcised, and $X_{ik}^{(1)}$ = "Female" if they are female (three levels are included as to not exclude females)
- $X_{ik}^{(2)} = 1$ if individual i in cluster k received HTC at enrollment or thereafter, and $X_{ik}^{(2)} = 0$ otherwise

$X_{ik}^{(12)}$ denotes whether individual i in cluster k received the entire treatment

- For males in BCPP, $X_{ik}^{(12)} = X_{ik}^{(1)} \times X_{ik}^{(2)} = 1$ if they received both MC and HTC, $X_{ik}^{(12)} = 0$ otherwise
- For females in BCPP, $X_{ik}^{(12)} = X_{ik}^{(2)} = 1$ if they received HTC, $X_{ik}^{(12)} = 0$ otherwise

$Z_k^{(1)}$, $Z_k^{(2)}$ is the proportion of indivividuals in village k who received the first component and second component of the treatment, respectively

- For males in BCPP, $Z_k^{(1)} = \sum_{i=1}^{m_k^{(\text{male})}} \frac{X_{ik}^{(1)}}{m_k^{(\text{male})}}$ is the proportion of males in village k who are circumcised before or during the study
- For all individuals in BCPP, $Z_k^{(2)} = \sum_{i=1}^{m_k} \frac{X_{ik}^{(2)}}{m_k}$ is the proportion of all individuals in village k who received HTC

$Z_{ik}^{(12)}$ is the proportion of individuals who received the full treatment

- For males in BCPP, $Z_{ik}^{(12)} = \sum_{i=1}^{m_k^{(\text{male})}} \frac{X_{ik}^{(1)} \times X_{ik}^{(2)}}{m_k^{(\text{male})}}$ is the proportion of males who are both circumcised and received HTC
- For females in BCPP, $Z_{ik}^{(12)} = Z_{ik}^{(2)} = \sum_{i=1}^{m_k^{(\text{female})}} \frac{X_{ik}^{(2)}}{m_k^{(\text{female})}}$ is the proprtion of females who received HTC

$\mathbf{C}_{ik} = (C_{1k}^{(1)}, ..., C_{m_k k}^{(1)}, C_{1k}^{(2)}, ..., C_{m_k k}^{(2)})$ are the individual level covariates

$\mathbf{V}_k = (V_k^{(1)}, ..., V_k^{(v)})$ are the cluster-level covariates

BASELINE CHARACTERISTICS

Characteristics Before Exclusions

The original dataset has 13131 total individuals in the study; 6591 in the treatment group, and 6540 in the control arm.

Variable	Level	Control	Treatment	Overall	Missing (Control)	Missing (Treatment)	Missing (Overall)
Number of Individuals		6540	6591	13131			
Number of Clusters		15	15	15			
Mean Cluster Size		459	461	460			
Gender	Female	4162	4178	8340			
	Male	2378	2413	4791			
HIV Status at Start	HIV-infected	1771	1825	3596	267	254	521
	HIV-uninfected	4487	4487	8974			
	Refused HIV testing	15	25	40			
Outcome: Seroconvert	Began study HIV-infected	1771	1825	3596	477	507	984
	No	4202	4202	8404			
	Yes	90	57	147			
Treatment Component: Full	No	3956	4044	8000	621	581	1202
	Yes	1963	1966	3929			
Treatment Component: HTC	No	3902	4008	7910	267	254	521
	Yes	2371	2329	4700			
Treatment Component: MC	Began study circumcised	652	777	1429	437	386	823
	Female	4162	4178	8340			
	No	1063	915	1978			
	Yes	226	335	561			

Table 1: Characteristics by treatment group before exclusions

Table below displays the mean proportion, per cluster, of various characteristics, including mean proportion of HIV infected individuals per cluster at baseline, etc. These are calculated before any exclusions.

Variable	Control	Treatment
Proportion of HIV Infected in Cluster (Mean, SD)	0.28 (0.07)	0.28 (0.07)
Proportion of Males in Cluster (Mean, SD)	0.36 (0.02)	0.36 (0.03)
Proportion of Males Circumcised in Cluster (Mean, SD)	0.37 (0.06)	0.46 (0.07)
Proportion HTC in Cluster (Mean, SD)	0.37 (0.06)	0.36 (0.05)
Proportion Fully Treated in Cluster (Mean, SD)	0.3 (0.05)	0.3 (0.05)

Table 2: Cluster-level proportions by treatment group before exclusions

Characteristics After Exclusions

A total of 3636 individuals were excluded from the analysis dataset. This is because these individuals either began the study as HIV-positive (n = 3596), or refused HIV testing (n = 40).

Note that in our analyses, we evaluated whether the intervention reduced HIV incidence by modeling seroconversion among individuals who were HIV-negative at baseline (n = 8974). Although the analysis was restricted to this at-risk subset, all cluster-level characteristics (e.g., proportion HIV-positive at baseline, proportion of men circumcised, etc.) were calculated using the full study population. This approach ensures that the covariates reflect the overall context and implementation environment of each cluster, rather than being limited to the analytic subset.

Variable	Level	Control	Treatment	Overall	Missing (Control)	Missing (Treatment)	Missing (Overall)
Number of Individuals		1786	1850	3636			
Gender	Female	1315	1343	2658			
	Male	471	507	978			
HIV Status at Start	HIV-infected	1771	1825	3596	267	254	521
	Refused HIV testing	15	25	40			
Outcome: Seroconvert	Began study HIV-infected	1771	1825	3596	477	507	984
Treatment Component: Full	No	293	332	625	621	581	1202
	Yes	1221	1244	2465			
Treatment Component: HTC	No	293	332	625	267	254	521
	Yes	1493	1518	3011			
Treatment Component: MC	Began study circumcised	84	108	192	437	386	823
	Female	1315	1343	2658			
	No	83	88	171			
	Yes	32	37	69			

Table 3: Characteristics by treatment group of the excluded individuals

MODELING RESULTS

Within-Village Spillover

Setup

- In this analysis, we include:
  a. Everyone in the treatment group who DID NOT receive *any* part of the treatment (For $T_k = 1$, $X_{ik}^{(1)} =$ "Yes" or "Female" and $X_{ik}^{(2)} = 0$)
  b. Everyone in the control group who DID NOT receive *any* part of the treatment (For $T_k = 0$, $X_{ik}^{(1)} =$ "Yes" or "Female" and $X_{ik}^{(2)} = 0$)
- This setup will allow us to estimate
  a. Spillover Within Intervention Clusters
  b. Spillover Mediated by Male Circumcision
  c. Proportion of Within-Intervention Village Spillover Effect Mediated by Male Circumcision

Dataset

```
# Only include those in treatment group who DID NOT receive any part of the treatment
# Only include those in control group who DID NOT receive any part of treatment
modelDat_SpW <- modelDat %>%
  filter(X1_ik != "Yes", X2_ik == 0) # Exclude anyone who got any part of the treatment
```

Summary counts of individuals per treatment group

```
   X1_ik
T_k Female   No
  0   2071  926
  1   2128  791
```

```
   X2_ik
T_k    0
  0 2997
  1 2919
```

```
      X1_ik
Y_ik   Female   No
  0      3965 1591
  1        82    0
  <NA>    152  126
```

```
      X2_ik
Y_ik      0
  0    5556
  1      82
  <NA>  278
```

A. Total Within-Cluster Spillover Effect of Treatment Assignment

"SpW" denotes total spillover within intervention clusters. This compares participants in intervention villages who received neither relevant intervention component to people in the control villages (who also did not receive any part of the intervention component)

Then, under certain assumptions, the only way for an intervention village participant to have lower HIV risk is by association with others in the village with lower HIV risk because of their exposure to the intervention.

$$logit(Y_{ik}) = \beta_0^{\text{SpW}} + \beta_1^{\text{SpW}}(T_k)$$

Then exp $\left(\beta_1^{\text{SpW}}\right)$ is a within-village spillover OR, and estimates the causal effect of living in a CP village, despite receiving no components oneself, on the odds of seroconversion. This is total within-village spillover effect.

```
# Spillover Within Intervention Clusters
model_SpW <- glm(Y_ik ~ T_k,
                 family = binomial(link = 'logit'),
                 data = modelDat_SpW) # Exclude those who received full trt

model_SpW_summary <- summary(model_SpW) # Save model summary

exp_beta_SpW_0 <- exp(model_SpW_summary$coefficients[1,1]) # Intercept
exp_beta_SpW_1 <- exp(model_SpW_summary$coefficients[2,1]) # T_k Coefficient

#model_SpW_summary
#tidy(model_SpW, exponentiate = TRUE, conf.int = TRUE) # Print output

tidy_SpW <- tidy(model_SpW, conf.int = TRUE, exponentiate = TRUE) %>%
  dplyr::select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  rename(
    Term = term,
    OR = estimate,
    SE = std.error,
    `p-value` = p.value,
    `95% CI (lower)` = conf.low,
    `95% CI (upper)` = conf.high
  )
# tidy_SpW
```

Term	OR	SE	p-value	95% CI (lower)	95% CI (upper)
(Intercept)	0.017	0.144	0.000	0.013	0.023
T_k	0.696	0.227	0.110	0.442	1.080

Table 4: Spillover Within Intervention Clusters Model Output

Thus, among people who received none of the intervention components, those living in CP villages had 30% lower odds of HIV seroconversion than otherwise comparable untreated people in control villages. Since every individual in this analytic set is personally untreated, any difference in their HIV risk can only arise from indirect protection, and thus, 0.7 is interpreted as the within-village spillover effect of CP.

B. Within-Cluster Spillover Treatment Assignment Effect Not through Male Circumcision

"SpWR" denotes all the remaining spillover that affects one's outcome that exists when we block the mediated spillover path that exists through male circumcision.

$$logit(Y_{ik}) = \beta_0^{\text{SpWR}} + \beta_1^{\text{SpWR}}(T_k) + \beta_2^{\text{SpWR}}(Z_k^{(1)})$$

Here, $\exp\left(\beta_1^{\text{SpWR}}\right)$ compares untreated individuals in CP villages with untreated individuals in control villages after we hold the village's male-circumcision coverage fixed at the same value for both groups. So, it's the OR for the remaining within-village spillover - whatever protection (or risk) is left once the male-circumcision pathway has been accounted for.

```
# Spillover Mediated by Male Circumcision
model_SpWR <- glm(Y_ik ~ T_k + Z1_k,
                  family = binomial(link = 'logit'),
                  data = modelDat_SpW)

model_SpWR_summary <- summary(model_SpWR) # Save model summary

exp_beta_SpWR_0 <- exp(model_SpWR_summary$coefficients[1,1]) # Intercept
exp_beta_SpWR_1 <- exp(model_SpWR_summary$coefficients[2,1]) # T_k Coefficient
exp_beta_SpWR_2 <- exp(model_SpWR_summary$coefficients[3,1]) # Z1_k Coefficient

tidy_SpWR <- tidy(model_SpWR, conf.int = TRUE, exponentiate = TRUE) %>%
  dplyr::select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  rename(
    Term = term,
    OR = estimate,
    SE = std.error,
    `p-value` = p.value,
    `95% CI (lower)` = conf.low,
    `95% CI (upper)` = conf.high
  )
```

Term	OR	SE	p-value	95% CI (lower)	95% CI (upper)
(Intercept)	0.054	0.683	0.000	0.014	0.209
T_k	0.895	0.271	0.683	0.524	1.517
Z1_k	0.046	1.846	0.095	0.001	1.563

Table 5: Spillover Mediated by Male Circumcision Model Output

After we hold village circumcision coverage fixed, untreated residence of CP villages will still have a 10% lower odds of seroconversion than untreated residence of control villages. This is spillover that operates through pathways other than male-circumcision coverage (e.g. HTC uptake, general behavior change, program outreach).

Then, moving from a 0% to 100% male circumcised coverage in a village multiplies an untreated person's odds of seroconversion by 0.05. This means 95% lower odds of HIV acquisition for an untreated person when their village goes from zero to complete male-circumcision coverage.

C. Proportion of Within-Intervention Village Spillover Treatment Assignment Effect Mediated by Circumcision

Then, $\frac{\beta_1^{\text{SpW}} - \beta_1^{\text{SpWR}}}{\beta_1^{\text{SpW}}}$ is the proportion of within-intervention village spillover effect mediated by circumcision.

```
# Proportion of within-intervention village spillover effect mediated by MC
(log(exp_beta_SpW_1) - log(exp_beta_SpWR_1)) / log(exp_beta_SpW_1)
```

```
[1] 0.6950287
```

Thus, about 70% of within-village spillover protection experienced by untreated people in CP villages is explained by the higher male-circumcision coverage in those villages. The reamining spillover benefit must come through other village-level channels (e.g. HTC uptake, community health behavior change, program outreach, etc.)

Individual Effects

Setup

- In this analysis, we include:
    a. All males in the treatment group
    b. All males in the control group
- In this analysis, we will estimate the effects of the intervention assignment on the outcome. In the mediation model, we will account for if they actually received the circumcision component or not.
- This setup will allow us to estimate
    a. Direct Effects of Treatment Assignment
    b. Mediated Effects of Circumcision
    c. Proportion of Individual Effect due to Circumcision

```
# Alternative to fix data availability
  # Include only those who were circumcised in the treatment
  # Include everyone in the control
modelDat_Ind <- modelDat %>%
  filter(C1_ik == 1)
```

Summary counts of individuals per treatment group

```
   X1_ik
T_k  No Yes <NA>
  0 926 717  146
  1 791 908   94
```

```
      X1_ik
Y_ik     No  Yes <NA>
  0    1591 1528  205
  1       0    9   19
  <NA>  126   88   16
```

```
   Y_ik
T_k    0    1 <NA>
  0 1659   20  110
  1 1665    8  120
```

D. Total Individual Effect of Treatment Assignment

"Ind" denotes individual effects, i.e. effects of a male's own treatment assignment on their own outcome. Here, we also block the spillover that exists through the proportion circumcised and proportion who received HTC in the cluster by controlling for it in the model.

$$logit(Y_{ik}) = \beta_0^{\text{Ind}} + \beta_1^{\text{Ind}}(T_k) + \beta_2^{\text{Ind}}(Z_k^{(1)}) + \beta_3^{\text{Ind}}(Z_k^{(2)})$$

The total individual effect of a male's own treatment assignment on their own outcome is β_1^{Ind}, and the corresponding OR is $\exp\left(\beta_1^{\text{Ind}}\right)$.

```
# Individual Effects
model_Ind <- glm(Y_ik ~ T_k + Z1_k + Z2_k,
                 family = binomial(link = 'logit'),
                 data = modelDat_Ind)

model_Ind_summary <- summary(model_Ind) # Save model summary

exp_beta_Ind_0 <- exp(model_Ind_summary$coefficients[1,1]) # Intercept
exp_beta_Ind_1 <- exp(model_Ind_summary$coefficients[2,1]) # T_k Coefficient,
                                                           # total Ind effect
exp_beta_Ind_3 <- exp(model_Ind_summary$coefficients[3,1]) # Z1 coeff
exp_beta_Ind_4 <- exp(model_Ind_summary$coefficients[4,1]) # Z2 coeff

#model_Ind_summary # Print summary
#tidy(model_Ind, exponentiate = TRUE, conf.int = TRUE) # Print output

tidy_Ind <- tidy(model_Ind, conf.int = TRUE, exponentiate = TRUE) %>%
  dplyr::select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  rename(
    Term = term,
    OR = estimate,
    SE = std.error,
    `p-value` = p.value,
    `95% CI (lower)` = conf.low,
    `95% CI (upper)` = conf.high
  )
#tidy_Ind
```

Since $\exp\left(\beta_1^{\text{Ind}}\right) = 0.47$, then the odds of seroconverting for a man in a CP village are 0.47 $\times$ the odds for a man in a control village. This is a 53% reduction in the odds. Thus, among men, being in a village randomized to the CP package is associated with substantially lower odds of acquiring HIV during the study.

E. Individual Direct Effects of Treatment Assignment

"IndD" denotes individual direct effects, i.e. effects of a male's own treatment assignment on their own outcome, blocking the path through the "mediator" (circumcision treatment component) by controlling for it in the model. Here, we also block the spillover that exists through the proportion circumcised and proportion who received HTC in the cluster by controlling for it in the model.

$$logit(Y_{ik}) = \beta_0^{\text{IndD}} + \beta_1^{\text{IndD}}(T_k) + \beta_2^{\text{IndD}}(X_{ik}^{(1)}) + \beta_3^{\text{IndD}}(Z_k^{(1)}) + \beta_4^{\text{IndD}}(Z_k^{(2)})$$

The individual direct effect of a male's treatment assignment on their outcome is thus β_1^{IndD}, since we are blocking the pathway from the treatment assignment to the outcome that goes through the circumcision component by controlling for it in the model.

The odds ratio of comparing CP-assignemnt vs. control for a male whose own circumcision status is held fixed is exp $\left(\beta_1^{\text{IndD}}(T_k)\right)$.

The odds ratio comparing circumcised vs. not, given village assignment is exp $\left(\beta_2^{\text{IndD}}(X_{ik}^{(1)})\right)$.

```
   Y_ik
T_k    0    1 <NA>
  0 1659   20  110
  1 1665    8  120
```

```
      Y_ik
X1_ik     0    1 <NA>
  No   1591    0  126
  Yes  1528    9   88
  <NA>  205   19   16
```

```
   X1_ik
T_k  No Yes <NA>
  0 926 717  146
  1 791 908   94
```

Because every uncircumcised man in our dataset had zero HIV-seroconversions, the data display complete separation. In other words, one predictor level ("No circumcision") perfectly predicts the outcome ("did not seroconvert"). Ordinary or exact logistic regression depends on being able to compare many hypothetical re-arrangements of the data in which the pattern is not perfect; when a column or row is all 0's or all 1's, those hypothetical tables don't exist. As a result, the likelihood tries to push the corresponding coefficient toward negative infinity, and the exact algorithm has nothing to sum over, so it simply cannot return an estimate.

Firth's bias-reduced logistic regression fixes the problem by adding a small penalty to the likelihood. The penalty keeps the estimates finite even when a predictor perfectly separates the outcome, and it also reduces small-sample bias. In practice, Firth's method produces odds-ratio estimates and profile-likelihood confidence intervals that coincide with exact (conditional) results whenever exact logistic can run, and it remains reliable when exact logistic cannot. Therefore we used the Firth-penalised model for the individual-direct-effect analysis; it is the standard remedy for separation and is widely accepted in epidemiology and biostatistics.

```
# Using Firth's model instead:
# Helpful when we have complete or quasi-complete separation
# It applies a penalty to the likelihood function that
# reduces small-sample bias in MLE
# Naturally prevents infinite estimates when separation occurs
# It uses Jeffreys invariant prior to modify the score function

model_IndD_firth <- logistf(Y_ik ~ T_k + X1_ik + Z1_k + Z2_k,
                            data = modelDat_Ind)

exp_beta_IndD_0_firth <- exp(model_IndD_firth$coefficients[[1]]) # Intercept
exp_beta_IndD_1_firth <- exp(model_IndD_firth$coefficients[[2]]) # T_k Coefficient,
                                                  # direct individual effect
exp_beta_IndD_2_firth <- exp(model_IndD_firth$coefficients[[3]]) # X1_ik Coefficient

exp_beta_IndD_3_firth <- exp(model_IndD_firth$coefficients[[4]]) # Z1_k Coefficient
```

```r
exp_beta_IndD_4_firth <- exp(model_IndD_firth$coefficients[[5]]) # Z2_k Coefficient

# Pull directly from the model
coef_table_firth <- model_IndD_firth$coefficients      # log-odds estimates (named vector)
se_table_firth   <- sqrt(diag(model_IndD_firth$var))   # standard errors
pval_table_firth <- model_IndD_firth$prob              # p-values
ci_lower_firth   <- model_IndD_firth$ci.lower          # lower bound on log-odds
ci_upper_firth   <- model_IndD_firth$ci.upper          # upper bound on log-odds

tidy_IndD_firth <- tibble(
  Term = names(coef_table_firth),
  OR = exp(coef_table_firth),
  SE = se_table_firth,
  `p-value` = pval_table_firth,
  `95% CI (lower)` = exp(ci_lower_firth),
  `95% CI (upper)` = exp(ci_upper_firth)
)
#tidy_IndM_firth
```

Holding a male's own circumcision status fixed, being in a village randomized to CP is associated with about 56% reduction in the odds of seroconversion compared with control villages. This is the individual controlled direct effect of assignment.

After accounting for village assignment, the model estimates that men who were circumcised have far higher observed odds of seroconversion than uncircumcised men (OR is 24.13). This very large odds-ratio arises because no seroconversions occurred among the uncircumcised group (complete separation); the Firth penalty keeps the estimate finite but it remains unstable and imprecise. Hence, the direction and size of this effect should be interpreted with caution - it is driven by sparse data rather than clear evidence that circumcision increases risk.

F. Indirect Individual Effect of Treatment Assignment

Then, the controlled indirect OR can be calculated as $\exp\left(\beta_1^{\text{Ind}} - \beta_1^{\text{IndD}}\right)$

```
exp(log(exp_beta_Ind_1) - log(exp_beta_IndD_1_firth)) # Controlled indirect individual effect
```

```
## [1] 1.064103
```

This captures the cluster assignment's effect on a man's seroconversion odds that operates through his own circumcision status (while all other pathways are held constant).

An odds-ratio of 1.06 means that, after we remove the "direct" pathway (OR 0.44), the remaining pathway that goes via a man's own circumcision increases his odds of seroconversion by roughly 6.41.

Overall, among men, the component of CP assignment that works through circumcision is associated with higher observed odds of HIV acquisition, offsetting some of the strong direct protection of being in a CP village.

Some caveats: - Since no uncircumcised men seroconverted, the circumcision coefficient (and therefore this ratio) is based on very little information and may be highly imprecise.

```
- The sign reversal (indirect pathway harmful while direct pathway protective) likely reflects residual confou

- So 1.54 tells us that, within the current model, the circumcision pathway moves the total effect closer to t
```

G. Proportion of Individual Effect of Treatment Assignment Mediated by Circumcision

The proportion of total individual effect of a male's treatment assignment, mediated by him receiving circumcision is $\frac{\beta_1^{\text{Ind}} - \beta_1^{\text{IndD}}}{\beta_1^{\text{Ind}}}$

```
(log(exp_beta_Ind_1) - log(exp_beta_IndD_1_firth))/(log(exp_beta_Ind_1))
```

```
## [1] -0.0828872
```

From this value, it seems that the mediator (one's own circumcision) works in the opposite direction to the total effect. Circumcision uptake appears to offset about 47% of the benefit that CP villages otherwise confer.

Overall Effects

H. Overall Intervention Village Effect

The overall effect of being in an intervention village can be calculated by just fitting the following model on the overall dataset of HIV-negative individuals (at the start of the study), without controlling for any other causal pathways.

$$logit(Y_{ik}) = \beta_0^{\text{Overall}} + \beta_1^{\text{Overall}}(T_k)$$

```
# Overall Effect of being in an Intervention Cluster
model_overall <- glm(Y_ik ~ T_k,
                     family = binomial(link = 'logit'),
                     data = modelDat) # Everyone

model_overall_summary <- summary(model_overall) # Save model summary

exp_beta_overall_0 <- exp(model_overall_summary$coefficients[1,1]) # Intercept
exp_beta_overall_1 <- exp(model_overall_summary$coefficients[2,1]) # T_k Coefficient

tidy_overall <- tidy(model_overall, conf.int = TRUE, exponentiate = TRUE) %>%
  dplyr::select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  rename(
    Term = term,
    OR = estimate,
    SE = std.error,
    `p-value` = p.value,
    `95% CI (lower)` = conf.low,
    `95% CI (upper)` = conf.high
  )
```

Term	OR	SE	p-value	95% CI (lower)	95% CI (upper)
(Intercept)	0.021	0.107	0.000	0.017	0.026
T_k	0.633	0.171	0.007	0.451	0.882

Table 6: Overall Effect of being in an Intervention Cluster

Thus, the OR is 0.63, meaning that for HIV-negative individuals at baseline, living in an intervention village is associated with a 37% reduction in the odds of seroconversion during follow-up compared with living in a control village. This single odds ratio blends every causal pathway, thus resulting in an overall effect. It is the total impact of the intervention environment on an average resident.

I. Proportion of total effect mediated by male circumcision The proportion of total effect mediated by male circumcision is

$$\frac{[(\beta_1^{\text{Ind}} - \beta_1^{\text{IndM}}) + (\beta_1^{\text{SpW}} - \beta_1^{\text{SpWM}})]}{(\beta_1^{\text{Ind}} + \beta_1^{\text{SpW}})}$$

```
# Proportion of total effect mediated by male circumcision
#((beta_Ind_1 - beta_IndM_1) + (beta_SpW_1 - beta_SpWM_1))/(beta_Ind_1 + beta_SpW_1)
# Proportion of total effect mediated by male circumcision firth model
exp(((log(exp_beta_Ind_1) - log(exp_beta_IndD_1_firth)) + (log(exp_beta_SpW_1) - log(exp_beta_SpWR_1)))/(log(e
```

```
## [1] 1.186343
```

J. Proportion of intervention village total effect due to spillover within intervention villages

The proportion of intervention village total effect due to spillover within intervention villages is

$$\frac{\beta_1^{\text{SpW}}}{(\beta_1^{\text{Ind}} - \beta_1^{\text{SpW}})}$$

```r
# Proportion of intervention village total effect due to spillover within intervention villages
log(exp_beta_SpW_1)/(log(exp_beta_Ind_1) + log(exp_beta_SpW_1))
```

```
## [1] 0.3262084
```

Note to Donna

I'm worried that this approach doesn't let us decompose the effects - if we want to be able to decompose the overall effect using estimands from the individual models and spillover models, we need to use the same population for all models, so that the models are truly nested.

Here's a summary of what we've done:

Model Name	Model equation	Analytic sample	Extra covariates	What β_1 estimates
Overall	$\text{logit}(Y_{ik}) = \beta_0 + \beta_1 T_k$	All HIV-negative participants	none	Total impact of CP village assignment (own uptake + *all* spillover)
Ind	$\text{logit}(Y_{ik}) = \beta_0 + \beta_1 T_k + \beta_2 Z_k^{(1)} + \beta_3 Z_k^{(2)}$	Males Only	village MC & HTC coverage	Assignment effect after measured spillover is held fixed (still contains own-uptake + unmeasured spillover)
IndD	$\text{logit}(Y_{ik}) = \beta_0 + \beta_1 T_k + \beta_2 X_{ik}^{(1)} + \beta_3 Z_k^{(1)} + \beta_4 Z_k^{(2)}$	Males Only	own MC $X_{ik}^{(1)}$ + Z's	Controlled-direct effect (paths via own MC **and** measured spillover blocked)
SpW	$\text{logit}(Y_{ik}) = \beta_0 + \beta_1 T_k$	Untreated Individuals	none	Total spillover for untreated people
SpWR	$\text{logit}(Y_{ik}) = \beta_0 + \beta_1 T_k + \beta_2 Z_k^{(1)}$	Untreated Individuals	village MC coverage	Remaining spillover after MC-coverage path is held fixed

And here's a proposal of how to maybe redo this so that the models are nested, and thus we can get valid results for important estimands such as "proportion of the total effect explained by spillover".

Steps:

1. Work on a single dataset (e.g. all HIV-negative participants at start of study)
2. Fit nested lgoistic models so that each new model adds only one class of pathways:

Label	Model equation	β_1 contains
A. Total effect	$\text{logit}\big(P(Y_{ik}=1)\big) = \beta_0 + \beta_1 T_k$	Own uptake + **all** spillover
B. Minus own uptake	$\text{logit}\big(P(Y_{ik}=1)\big) = \beta_0 + \beta_1 T_k + \beta_2 X_{ik}^{(1)} + \beta_3 X_{ik}^{(2)}$	**All spillover**
C. Minus measured spillover	$\text{logit}\big(P(Y_{ik}=1)\big) = \beta_0 + \beta_1 T_k + \beta_2 X_{ik}^{(1)} + \beta_3 X_{ik}^{(2)} + \beta_4 Z_k^{(1)} + \beta_5 Z_k^{(2)}$	Residual (*unmeasured*) spillover

3. On the log-odds scale, the proportion of the total effect explained by spillover is then

$$\frac{\beta_1^A - \beta_1^B}{\beta_1^A}$$

4. All decompositions:
   - Own uptake contribution is $\beta_1^A - \beta_1^B$
   - Measured spillover = $\beta_1^B - \beta_1^C$
   - Unmeasured spillover = β_1^C
   - These three pieces add up exactly: $\beta_1^A = (\beta_1^A - \beta_1^B) + (\beta_1^B - \beta_1^C) + \beta_1^C$

All quantities come from the same population, so the algebra and interpretation is valid.