BCPP Updated Analysis

Melody Owen

2025-08-01

Contents

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: Voluntary Medical 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.

Terminology

Individual Effects: Refers to an effect of one's own input on their own outcome

Spillover Effects: Refers to an effect of others' inputs on one's outcome; "within-cluster spillover" refers to spillover between individuals in the same cluster

Direct Effects: Refers to an effect pathway that links directly from an input to an output with nothing else on the pathway

Indirect Effects: Refers to an effect pathway that links indirectly from an input to an output through a node (mediator) on the pathway

With these, we can define the following:

Individual Effects

- **Individual Direct Effect**: Path from one's own input to their own outcome with no other nodes on the pathway
- **Individual Indirect Effect**: Path from one's own input to their own outcome through a node (mediator) on the pathway
- **Total Individual Effect**: Total effect of one's input on their outcome through all pathways (Individual Direct + Individual Indirect)

Spillover Effects

- **Spillover Direct Effect**: Path from others' inputs to one's outcome with no other nodes on the pathway
- **Spillover Indirect Effect**: Path from others' inputs to one's outcome through a node (mediator) on the pathway
- **Total Spillover Effect**: Total effect of other's inputs on one's outcome through all pathways (Spillover Direct + Spillover Indirect)

Overall Effects

- **Overall Effect**: Total Individual Effect + Total Spillover Effect

Questions of Interest

1. What is the direct individual effect of the CP intervention on HIV incidence?
2. To what extent is the total individual effect mediated by Voluntary Male Medical Circumcision (VMMC)?
3. What is the direct spillover effect of the CP intervention on HIV incidence?
4. To what extent is the total spillover effect mediated by VMMC?
5. What is the overall effect of CP on HIV incidence? ("overall" includes both spillover and individual totals)
6. To what extent is the overall effect mediated by VMMC?

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.
- $X_{ik}^{(1)}$ denotes any type of male circumcision (MC); $X_{ik}^{(1)}$ = "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)
- $X_{ik}^{(1,\text{VMMC})}$ denotes voluntary medical male circumcision (VMMC); $X_{ik}^{(1)}$ = "Yes" if individual i in cluster k was circumcised, $X_{ik}^{(1)}$ = "No" if they are male and not circumcised or circumcised before the study, 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 individuals 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 males in BCPP, $Z_k^{(1,\text{VMMC})} = \sum_{i=1}^{m_k^{(\text{male})}} \frac{X_{ik}^{(1,\text{VMMC})}}{m_k^{(\text{male})}}$ is the proportion of males in village k who have received VMMC 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
Number of Individuals		6540	6591	13131			
Number of Clusters		15	15	30			
Mean Cluster Size		459	461	460			
Gender	Male	2378 (36%)	2413 (37%)	4791 (36%)			
	Female	4162 (64%)	4178 (63%)	8340 (64%)			
HIV Status at Start	HIV-uninfected	4487 (72%)	4487 (71%)	8974 (71%)	267 (4%)	254 (4%)	521 (4%)
	HIV-infected	1771 (28%)	1825 (29%)	3596 (29%)			
	Refused HIV testing	15 (0%)	25 (0%)	40 (0%)			
Treatment Component: MC	Yes	548 (9%)	775 (12%)	1323 (11%)	437 (7%)	386 (6%)	823 (6%)
	No	1063 (17%)	915 (15%)	1978 (16%)			
	Began study circumcised	330 (5%)	337 (5%)	667 (5%)			
	Female	4162 (68%)	4178 (67%)	8340 (68%)			
Treatment Component: VMMC	Yes	548 (9%)	775 (12%)	1323 (11%)	437 (7%)	386 (6%)	823 (6%)
	No	1393 (23%)	1252 (20%)	2645 (21%)			
	Female	4162 (68%)	4178 (67%)	8340 (68%)			
Treatment Component: HTC	Yes	2371 (38%)	2329 (37%)	4700 (37%)	267 (4%)	254 (4%)	521 (4%)
	No	3902 (62%)	4008 (63%)	7910 (63%)			
Treatment Component: Full	Yes	1963 (33%)	1966 (33%)	3929 (33%)	621 (9%)	581 (9%)	1202 (9%
	No	3956 (67%)	4044 (67%)	8000 (67%)			
Outcome: HIV Seroconversion (3-year period)	Yes	90 (1%)	57 (1%)	147 (1%)	477 (7%)	507 (8%)	984 (7%)
	No	4202 (69%)	4202 (69%)	8404 (69%)			
	Began study HIV-infected	1771 (29%)	1825 (30%)	3596 (30%)			

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. Note that for the proportion of males circumcised in a given cluster, this includes both circumcision that occurred during and before the study.

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 of Males VMMC in Cluster (Mean, SD)	0.23 (0.1)	0.31 (0.13)
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 4580 individuals were excluded from the analysis dataset. This is because these individuals either began the study as HIV-positive (n = 3596), refused HIV testing (n = 40), or had a missing value (n = 521). Or, if they began the study HIV-negative, if they had a missing value for seroconversion (the outcome), they were excluded (n = 423).

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 = 8551). 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.

The following table shows the baseline characteristics of the new dataset that excludes these individuals (n = 8551).

Variable	Level	Control	Treatment	Overall	Missing (Control)	Missing (Treatment)	Missin
Number of Individuals		4292	4259	8551			
Number of Clusters		15	15	30			
Mean Cluster Size		460	465	462.5			
Gender	Male	1679 (39%)	1673 (39%)	3352 (39%)			
	Female	2613 (61%)	2586 (61%)	5199 (61%)			
HIV Status at Start	HIV-uninfected	4292 (100%)	4259 (100%)	8551 (100%)			
Treatment Component: MC	Yes	402 (10%)	588 (14%)	990 (12%)	138 (2%)	86 (1%)	224 (2
	No	868 (21%)	723 (17%)	1591 (19%)			
	Began study circumcised	271 (7%)	276 (7%)	547 (7%)			
	Female	2613 (63%)	2586 (62%)	5199 (62%)			
Treatment Component: VMMC	Yes	402 (10%)	588 (14%)	990 (12%)	138 (2%)	86 (1%)	224 (2
	No	1139 (27%)	999 (24%)	2138 (26%)			
	Female	2613 (63%)	2586 (62%)	5199 (62%)			
Treatment Component: HTC	Yes	844 (20%)	771 (18%)	1615 (19%)			
	No	3448 (80%)	3488 (82%)	6936 (81%)			
Treatment Component: Full	Yes	716 (17%)	690 (17%)	1406 (17%)	128 (2%)	81 (1%)	209 (2
	No	3448 (83%)	3488 (83%)	6936 (83%)			
Outcome: HIV Seroconversion (3-year period)	Yes	90 (2%)	57 (1%)	147 (2%)			
	No	4202 (98%)	4202 (99%)	8404 (98%)			

Table 3: Characteristics by treatment group after exclusions

MODELING RESULTS

Overall Effects

A. 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, n = 8551), without controlling for any other causal pathways.

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

```
# Overall Effects of T_k on Y_ik

# Model not accounting for clustering
model_overall <- glm(Y_ik ~ T_k,
                     family = binomial(link = 'logit'),
                     data = modelDat) # Everyone

# Model accounting for clustering using GLMM
model_overall_glmm <- glmer(Y_ik ~ T_k + (1|cluster_id),# Uses exchangeable
                            data = modelDat,
                            family = binomial(link = "logit"))

# Model accounting for clustering using GEE
model_overall_gee <- geeglm(Y_ik ~ T_k,
                            family = binomial(link = "logit"),
                            id = cluster_id,
                            data = modelDat,
                            corstr = "exchangeable") # working correlation
```

Model	Term	OR [95% CI]	p-value	ICC
GLM	T_k	0.633 [0.451, 0.882]	0.01	
GLMM	T_k	0.632 [0.432, 0.924]	0.02	0.02
GEE	T_k	0.63 [0.441, 0.899]	0.01	0.00

Table 4: Overall Effect of Treatment Assignment on HIV Model Output

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.

Within-Village Spillover

Setup

- In this analysis, we include everyone in the study (who began the study HIV-negative) who DID NOT receive any part of the treatment.
- This setup will allow us to estimate
  a. Total Within-Cluster Spillover Effect of the Intervention
  b. Within-Cluster Spillover of the Intervention Effect Not through Voluntary Medical Male Circumcision
  c. Proportion of Within-Intervention Village Spillover Effect Mediated by Voluntary Medical 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_vmmc != "Yes", X2_ik == 0) # Exclude anyone who got any part of the treatment
```

The total sample size for this analysis is 6096, meaning that 2455 individuals are excluded.

Data Characteristics

Variable	Level	Control	Treatment	Overall
Number of Individuals		3093	3003	6096
Number of Clusters		15	15	30
Mean Cluster Size		458	464	461
Gender	Male	1089 (35%)	960 (32%)	2049 (34%)
	Female	2004 (65%)	2043 (68%)	4047 (66%)
HIV Status at Start	HIV-uninfected	3093 (100%)	3003 (100%)	6096 (100%)
Treatment Component: MC	No	868 (28%)	723 (24%)	1591 (26%)
	Began study circumcised	221 (7%)	237 (8%)	458 (8%)
	Female	2004 (65%)	2043 (68%)	4047 (66%)
Treatment Component: VMMC	No	1089 (35%)	960 (32%)	2049 (34%)
	Female	2004 (65%)	2043 (68%)	4047 (66%)
Treatment Component: HTC	No	3093 (100%)	3003 (100%)	6096 (100%)
Treatment Component: Full	No	3093 (100%)	3003 (100%)	6096 (100%)
Outcome: HIV Seroconversion (3-year period)	Yes	51 (2%)	34 (1%)	85 (1%)
	No	3042 (98%)	2969 (99%)	6011 (99%)

Table 5: Characteristics of Spillover Effects Analysis Data

B. Total Within-Cluster Spillover Effect of the Intervention

"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.

```
# Total Within-Cluster Spillover Effect of the Intervention Model

# Model not accounting for clustering
model_SpW <- glm(Y_ik ~ T_k,
                 family = binomial(link = 'logit'),
                 data = modelDat_SpW) # Exclude those who received full trt

# Model accounting for clustering using GLMM
model_SpW_glmm <- glmer(Y_ik ~ T_k + (1|cluster_id), # Uses exchangeable structure
                        data = modelDat_SpW,
                        family = binomial(link = "logit"))

# Model accounting for clustering using GEE
model_SpW_gee <- geeglm(Y_ik ~ T_k,
                        family = binomial(link = "logit"), # logit link
                        id = cluster_id, # clustering variable
                        data = modelDat_SpW,
                        corstr  = "exchangeable") # working correlation
```

Model	Term	OR [95% CI]	p-value	ICC
GLM	T_k	0.683 [0.438, 1.053]	0.09	
GLMM	T_k	0.68 [0.429, 1.079]	0.10	0.01
GEE	T_k	0.681 [0.436, 1.062]	0.09	0.00

Table 6: Spillover Within Intervention Clusters Model Output

Thus, among people who received none of the intervention components, those living in CP villages had 32% 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, 0.68 is interpreted as the within-village spillover effect of CP.

C. Within-Cluster Spillover of the Intervention 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.

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

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.

```
# Within-Cluster Spillover of the Intervention Not Through MC

# Model not accounting for clustering
model_SpWR <- glm(Y_ik ~ T_k + Z1_k_vmmc,
                  family = binomial(link = 'logit'),
                  data = modelDat_SpW)

# Model accounting for clustering using GLMM
model_SpWR_glmm <- glmer(Y_ik ~ T_k + Z1_k_vmmc + (1|cluster_id), # Uses exchangeable
                         data = modelDat_SpW,
                         family = binomial(link = "logit"))

# Model accounting for clustering using GEE
model_SpWR_gee <- geeglm(Y_ik ~ T_k + Z1_k_vmmc,
                         family = binomial(link = "logit"),
                         id = cluster_id,
                         data = modelDat_SpW,
                         corstr = "exchangeable") # working correlation
```

Model	Term	OR [95% CI]	p-value	ICC
GLM	T_k	0.708 [0.443, 1.119]	0.14	
GLM	Z1_k_vmmc	0.62 [0.078, 4.301]	0.64	
GLMM	T_k	0.706 [0.434, 1.148]	0.16	0.01
GLMM	Z1_k_vmmc	0.621 [0.077, 5.021]	0.66	
GEE	T_k	0.706 [0.431, 1.157]	0.17	0.00
GEE	Z1_k_vmmc	0.622 [0.085, 4.575]	0.64	

Table 7: Spillover Not Due to Male Circumcision Model Output

After we hold village circumcision coverage fixed, untreated residence of CP villages will still have a 29% 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 0.62. This means 38% lower odds of HIV acquisition for an untreated person when their village goes from zero to complete male-circumcision coverage.

D. Proportion of Within-Intervention Village Spillover Effect Mediated by Circumcision

Then, the proportion of within-intervention village spillover effect mediated by circumcision is

$$\frac{\beta_1^{\text{SpW}} - \beta_1^{\text{SpWR}}}{\beta_1^{\text{SpW}}}$$

Essentially, this is the total spillover minus the spillover that exists except through the circumcision component, divided by spillover total.

Model	Proportion of Spillover Mediated by MC
GLM	0.10
GLMM	0.10
GEE	0.10

Table 8: Proportion of Within-Cluster Spillover Due to Male Circumcision

Thus, about 10% of within-village spillover protection experienced by untreated people in CP villages is explained by the higher male-circumcision coverage in those villages. The remaining 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 only males in the study.
- Here, 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

    d. Total Individual Effect of Treatment Assignment
    e. Individual Direct Effects of Treatment Assignment
    f. Indirect Individual Effect of Treatment Assignment
    g. Proportion of Individual Effect of Treatment Assignment Mediated by Circumcision

```
# Alternative to fix data availability
  # Include only those who were circumcised in the treatment
  # Include everyone in the control
modelDat_Ind <- modelDat %>%
  filter(gender == "Male") %>%
  filter(!is.na(X1_ik_vmmc))
```

The total sample size for this analysis is 3128, meaning that 5423 individuals are excluded (5199 females and 224 with missing data for VMMC).

Data Characteristics

Variable	Level	Control	Treatment	Overall
Number of Individuals		1541	1587	3128
Number of Clusters		15	15	30
Mean Cluster Size		461	465	463
Gender	Male	1541 (100%)	1587 (100%)	3128 (100%)
HIV Status at Start	HIV-uninfected	1541 (100%)	1587 (100%)	3128 (100%)
Treatment Component: MC	Yes	402 (26%)	588 (37%)	990 (32%)
	No	868 (56%)	723 (46%)	1591 (51%)
	Began study circumcised	271 (18%)	276 (17%)	547 (17%)
Treatment Component: VMMC	Yes	402 (26%)	588 (37%)	990 (32%)
	No	1139 (74%)	999 (63%)	2138 (68%)
Treatment Component: HTC	Yes	107 (7%)	147 (9%)	254 (8%)
	No	1434 (93%)	1440 (91%)	2874 (92%)
Treatment Component: Full	Yes	107 (7%)	147 (9%)	254 (8%)
	No	1434 (93%)	1440 (91%)	2874 (92%)
Outcome: HIV Seroconversion (3-year period)	Yes	7 (0%)	2 (0%)	9 (0%)
	No	1534 (100%)	1585 (100%)	3119 (100%)

Table 9: Characteristics of Individual Effects Analysis Data

The mediator model regresses the mediator on the exposure and confounders. Here, we block the spillover that exists through the proportion circumcised and proportion who received HTC in the cluster by controlling for it in the model (inclusion of $Z_k^{(1)}$ and $Z_k^{(2)}$ in the model).

E. Mediator Model for Individual Effect of Treatment Assignment

"IndM" denotes individual effects, i.e. effects of a male's own treatment assignment on their own outcome. Here, we use the product method to calculate the direct and indirect effects, with the outcome being Y_{ik}, treatment being T_k, and mediator being $X_{ik}^{(1,\text{VMMC})}$, whether or not a male has VMMC. The abbreviation "IndM" refers to the individual effect mediator model, shown below.

$$logit(X_{ik}^{(1,\text{VMMC})}) = \beta_0^{\text{IndM}} + \beta_1^{\text{IndM}}(T_k)$$

```
# Mediator Model for Individual Effect of Treatment Assignment

modelDat_Ind <- modelDat_Ind %>%
  mutate(X1_ik_vmmc = ifelse(X1_ik_vmmc == "Yes", 1,
                             ifelse(X1_ik_vmmc == "No", 0, NA)))

# Model not accounting for clustering
model_IndM <- glm(X1_ik_vmmc ~ T_k,
                  family = binomial(link = 'logit'),
                  data = modelDat_Ind)

# Model accounting for clustering using GLMM
model_IndM_glmm <- glmer(X1_ik_vmmc ~ T_k + (1|cluster_id), # Uses exchangeable
                         data = modelDat_Ind,
                         family = binomial(link = "logit"))

# Model accounting for clustering using GEE
model_IndM_gee <- geeglm(X1_ik_vmmc ~ T_k,
                         family = binomial(link = "logit"),
                         id = cluster_id,
                         data = modelDat_Ind,
                         corstr = "exchangeable") # working correlation
```

Model	Term	OR [95% CI]	p-value	ICC
GLM	T_k	1.668 [1.432, 1.944]	0.00	
GLMM	T_k	1.686 [1.047, 2.716]	0.03	0.10
GEE	T_k	1.634 [1.03, 2.591]	0.04	0.09

Table 10: Mediator Model for Individual Effects

F. Outcome Model for Individual Effect of Treatment Assignment

Now, we regress the outcome, Y_{ik} on the treatment assignment T_k and VMMC mediator $X_{ik}^{(1,\text{VMMC})}$. The abbreviation "IndO" refers to the individual effect outcome model, shown below.

$$logit(Y_{ik}) = \beta_0^{\text{IndO}} + \beta_1^{\text{IndO}}(T_k) + \beta_2^{\text{IndO}}(X_{ik}^{(1,\text{VMMC})})$$

```
# Outcome Model for Individual Effect of Treatment Assignment

# Model not accounting for clustering
model_IndO <- glm(Y_ik ~ T_k + X1_ik_vmmc,
                  family = binomial(link = 'logit'),
                  data = modelDat_Ind)

# Model accounting for clustering using GLMM
model_IndO_glmm <- glmer(Y_ik ~ T_k + X1_ik_vmmc + (1|cluster_id), # Uses exchangeable
                         data = modelDat_Ind,
                         family = binomial(link = "logit"))

# Model accounting for clustering using GEE
```

```r
model_IndO_gee <- geeglm(Y_ik ~ T_k + X1_ik_vmmc,
                         family = binomial(link = "logit"),
                         id = cluster_id,
                         data = modelDat_Ind,
                         corstr = "exchangeable") # working correlation
```

Model	Term	OR [95% CI]	p-value	ICC
GLM	T_k	0.226 [0.034, 0.948]	0.07	
GLM	X1_ik_vmmc	5.139 [1.342, 24.532]	0.02	
GLMM	T_k	0.216 [0.032, 1.46]	0.12	0.30
GLMM	X1_ik_vmmc	5.615 [1.354, 23.283]	0.02	0.30
GEE	T_k	0.226 [0.045, 1.144]	0.07	0.00
GEE	X1_ik_vmmc	5.139 [1.191, 22.181]	0.03	0.00

Table 11: Outcome Model for Individual Effects

G. Direct and Indirect Individual Effects of Treatment Assignment

The direct individual effect of treatment assignment on HIV seroconversion is β_1^{IndO}. The indirect individual effect of treatment assignment on HIV seroconversion is $\beta_1^{\text{IndM}} \times \beta_2^{\text{IndO}}$. Then, the total effect is the sum of these, namely $\beta_1^{\text{IndO}} + \beta_1^{\text{IndM}} \times \beta_2^{\text{IndO}}$.

Model	Direct Effect	Indirect Effect	Total Effect
GLM	0.23	2.31	0.52
GLMM	0.22	2.46	0.53
GEE	0.23	2.23	0.51

Table 12: Direct, Indirect, and Total Individual Effects of Treatment Assignment (Shown as ORs)

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

The proportion of the total individual effect that is mediated by circumcision can be calculated as:

$$\frac{\beta_1^{\text{IndM}} \times \beta_2^{\text{IndO}}}{\beta_1^{\text{IndO}} + \beta_1^{\text{IndM}} \times \beta_2^{\text{IndO}}}$$

This is the indirect effect (i.e. effect of treatment assignment through the MC mediator on the outcome) divided by the total effect of the treatment assignment on the outcome (both through and not through the mediator).

Model	Proportion Mediated
GLM	-1.29
GLMM	-1.43
GEE	-1.18

Table 13: Proportion of Individual Effect of Treatment Assignment Mediated by Circumcision

Calculating the Confidence Interval for Proportion Mediated Delta Method: First-order Taylor approximation for the variance of a smooth function $g(\hat{\theta})$ of estimators, where

$$\text{Var}\left\{g(\hat{\theta})\right\} \approx \nabla g(\hat{\theta})^T \hat{\text{Var}}(\hat{\theta}) \nabla g(\hat{\theta})$$

where ∇g is the gradient of g with respects to θ evaluated at the estimates.

To simplify notation, let $a = \beta_1^{\text{IndM}}$ (from the mediator logit model), $b_1 = \beta_1^{\text{IndO}}$ and $b_2 = \beta_2^{\text{IndO}}$ (from the outcome logit model). Then, NIE $= ab_2$, TE $= b_1 + ab_2$, and PM $= \frac{\text{NIE}}{\text{TE}} = \frac{ab_2}{b_1 + ab_2}$ (natural indirect effect, total effect, and proportion mediated, respectively). The gradient for PM with respects to (a, b_1, b_2) is

$$\nabla \text{PM} = \left(\frac{b_1 b_2}{(b_1 + ab_2)^2}, -\frac{ab_2}{(b_1 + ab_2)^2}, \frac{ab_1}{(b_1 + ab_2)^2} \right)$$

Then SE(PM) $\approx \sqrt{\nabla^T \Sigma \nabla}$, where Σ is the joint covariance matrix of (a, b_1, b_2). The mediator and outcome coefficients come from different models on the same data, so in principle, Σ has cross-covariances between a and (b_1, b_2). Those aren't available by default, but we can use bootstrapping. We can use bootstrapping to find an estimated joint covariance of the logit coefficients (a, b_1, b_2) to obtain $\hat{\Sigma} = \text{Cov}_{\text{boot}}(\hat{a}, \hat{b}_1, \hat{b}_2)$, and then apply

$$\text{Var}(\hat{\text{PM}}) \approx \nabla \text{PM}(\hat{\theta})^T \hat{\Sigma} \nabla \text{PM}(\hat{\theta})$$

More specifically, in each bootstrap replicate $b = 1, ..., B$, you refit both models on the same resampled clusters, giving the triplet $\hat{\theta}^{(b)} = (\hat{a}^{(b)}, \hat{b}_1^{(b)}, \hat{b}_2^{(b)})$. The joint covariance matrix $\hat{\Sigma}$ is just the sample covariance of those triplets:

$$\hat{\Sigma} = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \bar{\theta})(\hat{\theta}^{(b)} - \bar{\theta})^T, \bar{\theta} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}$$

The diagonals are the boot variances of a, b_1, and b_2. The off-diagonals (e.g. $\hat{Cov}(a, b_2)$) capture their dependence across models, because we estimated them from the same resamples. This way, we don't need to falsely assume independence between the mediator and outcome fits.

Below shows the CI for the PM using the delta method (and bootstrapping the joint covariance), and also the CI without using the delta method and purely just using the bootstrap. The percentile bootstrap CIs are extremely susceptible to sparse data, thus causing a wide interval. However, the delta-method CI with boot $\hat{\Sigma}$ reflects local uncertainty and preserves the true cross-model dependence, leading to narrower and easier to interpret CIs.

GLM CI for PM

```
set.seed(123)
res <- pm_boot_with_delta(modelDat_Ind, id = "cluster_id", B = 1000, verbose = TRUE)
```

```
## bootstrap: 100/1000
## bootstrap: 200/1000
## bootstrap: 300/1000
## bootstrap: 400/1000
## bootstrap: 500/1000
## bootstrap: 600/1000
## bootstrap: 700/1000
## bootstrap: 800/1000
## bootstrap: 900/1000
## bootstrap: 1000/1000
```

```
res$point_PM
```

```
## [1] -1.291073
```

```
res$ci_boot_PM
```

```
##      2.5%     97.5%
## -22.71376  32.77719
```

```
res$ci_delta_PM
```

```
## [1] -4.299618  1.717471
```

GLMM CI for PM

```
set.seed(123)
res_glmm <- pm_boot_with_delta_glmm(modelDat_Ind, id = "cluster_id", B = 300)
```

```
## bootstrap: 30/300
## bootstrap: 60/300
## bootstrap: 90/300
```

```
## bootstrap: 120/300
## bootstrap: 150/300
## bootstrap: 180/300
## bootstrap: 210/300
## bootstrap: 240/300
## bootstrap: 270/300
## bootstrap: 300/300
```

```
res_glmm$point_PM
```

```
## [1] -1.431248
```

```
res_glmm$ci_boot_PM
```

```
##      2.5%     97.5%
## -29.94983  20.48655
```

```
res_glmm$ci_delta_PM
```

```
## [1] -4.787820  1.925324
```

GEE CI for PM

```
set.seed(123)
res_gee <- pm_boot_with_delta_gee(modelDat_Ind, id = "cluster_id",
                                  B = 300, corstr = "exchangeable")
```

```
## bootstrap: 30/300
## bootstrap: 60/300
## bootstrap: 90/300
## bootstrap: 120/300
## bootstrap: 150/300
## bootstrap: 180/300
## bootstrap: 210/300
## bootstrap: 240/300
## bootstrap: 270/300
## bootstrap: 300/300
```

```
res_gee$point_PM
```

```
## [1] -1.17823
```

```
res_gee$ci_boot_PM
```

```
##      2.5%     97.5%
## -20.70932  31.54286
```

```
res_gee$ci_delta_PM
```

```
## [1] -3.727113  1.370652
```

Total Proportion Mediated

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

(Total Individual - Direct Individual) + (Total Spillover - Spillover Not Mediated by MC) / Overall (Or Total Individual + Total Spillover)

$$\frac{[(\beta_1^{\text{IndM}} \times \beta_2^{\text{IndO}}) + (\beta_1^{\text{SpW}} - \beta_1^{\text{SpWR}})]}{(\beta_1^{\text{Overall}})}$$

Model	Proportion of Overall Effect Mediated by Circumcision
GLM	-1.75
GLMM	-1.88
GEE	-1.66

Table 14: Proportion of Overall Effect of Treatment Assignment Mediated by Circumcision

CURRENT ISSUES AND NOTES

Ashley's email:

1. It would be helpful to write out the estimands for each effect. Are these the same as Tyler's paper? Does it matter that your mediator is part of the intervention package, while Tyler's is another covariate? https://pmc.ncbi.nlm.nih.gov/articles/PMC3753117/
2. Are the meditators at the individual level, group level or both? If group level, are you using an exposure mapping function?
3. Can the spillover effect itself be mediated? In the case of just two people in a cluster, I believe the spillover problem can be exactly described as meditation.
4. How are the other package components handled? Could treat as confounders, averaging over them.
5. For the OR =22, I would check the table of village assignment, VMMC exposure and outcome, probably a zero or small cell. Ke, any other ideas here?
6. Are you concerned about mediator confounding in this case? How is the correlation within cluster modeled?
7. I think the positivity assumption is OK - as you have a three level exposure (women, male cirm, male unicorn) and fairly sure there are no gender restrictions on the covariates., but the "woman" status is not an intervention like VMMC is.

Notes from Laura

1. what we are doing is first estimating the total effect (which is a sum of the direct and indirect effects) by regressing on T.
    a. Laura's response: If regressing on T, this is the overall (total) effect, that is the overall effect of being in an intervention cluster compared to control. It's total because it's regardless of mediation.
2. then we are adjusting for x1 (vmmc) to get the direct effect of CP (assuming perfect compliance) and we can get the indirect effect through vmmc through the usual method we have in the slides.
    a. This is the overall indirect effect, that is the effect of cluster assignment through individual VMMC.

the new suggestion allows for imperfect compliance and this uses X12 in place of T.

`a. What is the new suggestion?`

Other Notes

1. Circumcision that is done locally, not for medical purposes but for cultural purposes, is incomplete and may not be effective for preventing HIV. So at some point later in the analysis, it would be of interest to assess the effects of circumcision before the study started, with circumcision after.
2. Positivity assumption was mentioned as a possible issue because women can't have VMMC
3. Add a dag for mediation, and a dag for spillover - has anybody ever done this?
4. Do we have data on death? We can combine HIV and death as another outcome to have more events
5. There is a variable called hiv_status_time that gives HIV status by each visit, with status already positive, new positive, negative. Do you see this? And then there is a variable hiv_results_days that is days from enrollment to that test. That could give us 1 year incidence when/if we want that (for survival data analysis for example). These should add up to the number of cases in total.
6. If we want to create a combined variable, death or seroconversion, which could give us more cases and more power, if all or most of the deaths are due to HIV.
7. Time from enrollment to death is death_days. There is also a variable death_cause (can we take a look at this? It might be possible to delete deaths that are obviously not HIV related, such as accidents). And then there is another variable death_primary, which is the primary cause of death.

MC and VMMC

Evidence that VMMC and MC has no individual effect on HIV incidence and therefore cannot mediate the intervention

X1_ik	Y_ik = Yes	Y_ik = No
X1_ik = Yes	9	1528
X1_ik = No	0	1591
X1_ik = Missing	19	205

Table 15: Counts of HIV Seroconversion (Y) by Male Circumcision (X1)

X1_ik_vmmc	Y_ik = Yes	Y_ik = No
X1_ik_vmmc = Yes	6	984
X1_ik_vmmc = No	3	2135
X1_ik_vmmc = Missing	19	205

Table 16: Counts of HIV Seroconversion (Y) by VMMC (X1)

Model	Term	OR [95% CI]	p-value	ICC
GLM	X1_ik	37191028.878 [0, NA]	0.99	
GLMM	X1_ik	489227758283.171 [0, 6.21100157418884e+106]	0.81	0.46
GEE	X1_ik	Inf [Inf, Inf]	0.00	0.00

Table 17: Effect of MC on HIV Incidence

Fisher's exact test

```
mat <- matrix(c(9, 1528,   # first row
                0, 1591),  # second row
              nrow = 2, byrow = TRUE)
fisher.test(mat, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  mat
## p-value = 0.00165
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.050316      Inf
## sample estimates:
## odds ratio
##        Inf
```

```
vmmcmat <- matrix(c(6, 984,   # first row
                3, 2135),  # second row
              nrow = 2, byrow = TRUE)
fisher.test(vmmcmat, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  vmmcmat
## p-value = 0.03276
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.9240361 26.8562015
## sample estimates:
## odds ratio
##   4.337147
```

Evidence that village-level circumcision rates have a strong effect and mediate treatment effect in men.

Model	Term	OR [95% CI]	p-value	ICC
GLM	X1_ik_vmmc	4.339 [1.142, 20.596]	0.04	
GLMM	X1_ik_vmmc	5.243 [1.249, 22.016]	0.02	0.45
GEE	X1_ik_vmmc	4.538 [1.148, 17.94]	0.03	0.00

Table 18: Effect of VMMC on HIV Incidence

Mediator Model:

$$Z_k^{(1)} = \beta_0^{\mathrm{M}} + \beta_1^{\mathrm{M}}(T_k)$$

Model	Term	Beta [95% CI]	p-value
LM	T_k	0.089 [0.086, 0.092]	0.00
LMM	T_k	0.014 [0.011, 0.017]	0.00
GEE	T_k	0.095 [0.05, 0.14]	0.00

Table 19: Mediatior Model for Proportion VMMC

Outcome Model:

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

Model	Term	OR [95% CI]	Beta [95% CI]	p-value	ICC
GLM	T_k	0.791 [0.526, 1.184]	-0.234 [-0.643, 0.169]	0.26	
GLM	Z1_k	0.078 [0.005, 1.079]	-2.552 [-5.275, 0.076]	0.06	
GLMM	T_k	0.789 [0.51, 1.221]	-0.237 [-0.673, 0.2]	0.29	0.01
GLMM	Z1_k	0.085 [0.005, 1.47]	-2.471 [-5.327, 0.386]	0.09	0.01
GEE	T_k	0.787 [0.514, 1.207]	-0.239 [-0.666, 0.188]	0.27	0.00
GEE	Z1_k	0.081 [0.005, 1.447]	-2.51 [-5.389, 0.37]	0.09	0.00

Table 20: Outcome Model for Proportion VMMC

Model	Direct Effect	Indirect Effect	Total Effect	Proportion Mediated
LM/GLM	-0.23	-0.23	-0.46	0.49
LMM/GLMM	-0.24	-0.03	-0.27	0.13
GEE	-0.24	-0.24	-0.48	0.50

Table 21: Direct, Indirect, Total, and Proportion Mediated Proportion VMMC