Web Appendix: Technical details

Overview

We estimated the cumulative risk of infection under various exposure plans in which we intervened on both the exposure and mediators in a Monte Carlo simulation of 10,000 young women randomly sampled with replacement at baseline. The probabilities of covariates measured at baseline were estimated using their empirical distributions in the sampled data, and the probabilities of time-varying covariates and mediators were modeled using parametric regression models in the observed data, conditional on confounders in the observed data. Confounders were selected using directed acyclic graphs of the exposure-outcome relationship and mediator-outcome and relationships, based on existing literature. We compared our predicted risk under no intervention on the exposure or mediators (i.e., under the “natural course”19) with the observed data to assess the fit of the parametric models.19,24In addition, we tested for interaction between the exposure and each mediator by comparing risks from models with and without an interaction term and by using a likelihood ratio test. We were unable to include an interaction term due to sparse data. However, when using mediators and exposures at the same time point we found no evidence of interaction and it would be even less likely for an exposure to interact with the mediator at the following time point. We did include an interaction term between the mediator partner age difference and partner number to intervene independently and jointly on the mediators. Lastly, the probability of infection was estimated for each young woman at each follow up visit using a parametric regression model with exposure assignment and mediator set to what it would have been under the given exposure plan. We simulated the cohort with no loss to follow up. All girls had 5 visits corresponding to three observations (rows) in the dataset (Figure 1). In the observed data, the average number of visits with exposure ascertainment was 1.7 versus 3 in the simulated data. In the Monte Carlo simulation, row 1 includes exposure time 1, mediator time 2 and outcome time 3. Row 2 includes exposure time 2, mediator time 3 and outcome time 4. Row 3 includes exposure time 3, meditator time 5, outcome time 5.

The simulated cohort closely replicated the observed data with very minimal differences in population characteristics (Appendix table 1). The 10,000 simulated young women had 29,354 person-visits with 96.0% (28,198) attending 80% or more school days. At baseline in the observed data, 5.1% (N=105) had a partner five or more years older, 78.9% had zero partners in the last 12 months (N=1,626), 16.7% (N=344) had one partner and 4.5% (N=92) had two or more partners. At baseline in the simulated cohort, 5.2% (N=5,222) had a partner five or more years older, 78.3% had zero partners in the last 12 months (N=7,719), 17.1% (N=1,687) had one partner and 4.6% (N=454) had two or more partners. Cumulative incidence of HIV and HSV-2 estimated under no intervention on exposure or mediators (the “natural course”) was similar to the cumulative incidence of the outcomes in the observed data (Appendix Figure 1).

Details

Let patients be indexed from 𝑖 = 1, ... 2,086, K represent school attendance, L1 represent a vector of exposure-outcome confounders including covariates at baseline (age, intervention assignment, socioeconomic status (SES), orphan status, depression and anxiety) and time-varying confounders (prior HSV-2 or HIV status, prior alcohol use, and follow up visit). L2 represents time-varying mediator-outcome confounders of depression, anxiety and alcohol use. Z represents time-varying mediators partner age difference and partner number; and an interaction term between the mediators. We write the cumulative incidence at time t + 2 using the g-formula:

Equation 1:

K(t): School attendance

Y(t2): HIV and HSV-2

Z(t+1): Mediators partner age and partner number under the natural course (under some interventions the probability was set to 1)

L1(t): Confounders of K->Y

L2(t+1): Confounders of Z->Y

At t<2 L1(t-2)=L1(baseline)

At t<1 L2(t)=L2(baseline)

To estimate the cumulative incidence under each exposure plan from the observed data, we follow the steps below:

1. Fit parametric models for each component of the density in the observed data. In all models, categorical variables were modelled using indicator variables.
2. Fit a logistic model to estimate the conditional probability of each time-varying confounder of the exposure-outcome relationship (L1) for young woman i at time t. Time-varying exposure-outcome confounders include prior alcohol use, and prior HSV-2 status (or HIV status for HSV-2 outcome). V is a vector of time-fixed covariates (intervention arm at baseline, age at baseline, SES at baseline (quartiles), anxiety at baseline and depression at baseline) and time-varying follow-up visit number.
3. Fit a logistic model to estimate for the time-varying mediator outcome confounders (L2) for patient i at time t. Mediator-outcome confounders include alcohol use, depression and anxiety. Where V is a vector of time-fixed covariates (intervention arm at baseline, age at baseline, anxiety at baseline and depression at baseline) and time-varying follow-up visit number. The exposure of prior high attendance (K) was also included for all models. For the alcohol and anxiety mediators, depression at that time point was included. Additionally, anxiety at that time point was included for alcohol use.

1. Fit a logistic model to estimate the time-varying mediator of partner age difference (Z) for young woman i at time t. Where V is a vector of time-fixed covariates (intervention arm at baseline, and age at baseline) and time-varying follow up visit number. The exposure of high attendance (K) was also included as well as mediator-outcome confounders ( L2 is vector of mediator-outcome confounders (depression, anxiety and alcohol use).
2. Fit proportional odds model to estimate the time-varying mediator of partner number (Z) for young woman i at time t. G is the level of the mediator partner number (1 or 2). V is a vector of time-fixed covariates (intervention arm at baseline, and age at baseline) and time-varying follow-up visit number. The exposure of high attendance (K) was also included as well as mediator-outcome confounders ( L2 is vector of mediator-outcome confounders (depression, anxiety and alcohol use).
3. Fit logistic model to estimate the outcome of HIV or HSV-2 infection (Y) for young women i at time t. Where V is a vector of time-fixed covariates (intervention arm at baseline, age at baseline, SES at baseline, anxiety at baseline, depression at baseline, orphan status at baseline). Variables that are also included are: the exposure of high attendance K; L1 vector of time- varying confounders (prior alcohol use, prior HSV or HIV, follow up visit number) status; L2 is vector of mediator-outcome confounders (depression, anxiety and alcohol use); Z mediators partner number and partner age difference and an interaction between the mediators.



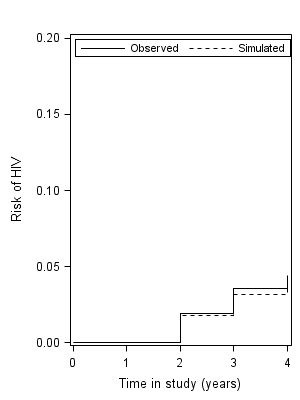
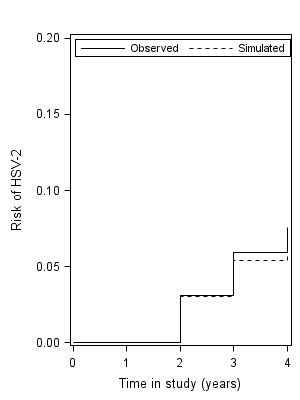
1. Fit a logistic model to estimate for the exposure of high versus low attendance in school (K) for patient i at time t. Where V is a vector of time-fixed covariates, intervention arm at baseline, age at baseline, SES at baseline (quartiles), anxiety at baseline and depression at baseline. L1 is a vector of time-varying covariates including follow-up visit number, alcohol use, and time-varying HSV-2 (or HIV) infection. The exposure was modelled under the natural course but was set for the various intervention scenarios.



1. Draw a large (N = 10,000) Monte Carlo sample from the observed patients at baseline with replacement.
2. In the Monte Carlo sample, we estimated the cumulative incidence under no intervention (as a check on the fit of the parametric models) and each dynamic treatment plan using the g-formula provided in equation 1.
   1. The distribution of L in the large Monte Carlo sample approximates the distribution of L in the observed data.
   2. Estimate time-varying confounders, exposure, time-varying mediator-outcome confounders, time-varying mediators and outcome using probabilities estimated above at each time point until infection or end of the time period.
   3. Increase the variable for follow-up visit by 1 at each iteration.
   4. Set treatment and mediators according to the exposure plans outlined in the manuscript.
3. Perform steps 1 through 4 in 200 bootstrap samples. The standard deviation of the 200 estimates can be used as the standard error of the point estimate.

To identify the controlled direct effect we assume 1) no unmeasured confounding of the exposure-outcome relationship, 2) no unmeasured confounding for the mediator-outcome relationship, and 3) that there is a hypothetical intervention on the mediator (partner age difference or number of partners) that could set the mediator to one of the values examined. [24,25] Setting partner age or number of partners to one level for all young women may be possible by creating environments like school, but is not practical as an intervention. However, we believe our estimates helps to illustratively answer the question of whether we should think about these behaviors as network factors that are important mediators and, if so, how we could intervene to change them.[40] In addition, we show that if young women had 50% fewer older partners and all had fewer partners, the effect of school attendance on incidence of infection would be much smaller.

**Appendix figure 1:** Observed versus simulated cumulative incidence of HIV and HSV-2 by time since study enrollment

**Appendix table 1:** Baseline characteristics of young women aged 13 to 20 without prevalent HIV infection and at least three follow-up visits in Agincourt, South Africa from March 2011 to December 2012

|  |  |  |
| --- | --- | --- |
|  | Observed (N=2,086)\*  N (%) | Simulated (N=10,000)  N (%) |
| Young women’s age at baseline (year) |  |  |
| Age 13-14 | 717 (34.4) | 3378 (33.8) |
| Age 15-16 | 913 (43.8) | 4423 (44.2) |
| Age 17-18 | 384 (18.4) | 1847 (18.5) |
| Age 18-20 | 72 (3.5) | 352 (3.5) |
| Household wealth |  |  |
| Low | 540 (25.9) | 2599 (26.0) |
| Middle to Low | 566 (27.2) | 2661 (26.7) |
| Middle | 489 (23.5) | 2316 (23.2) |
| High | 488 (23.4) | 2405 (24.1) |
| CCT randomization arm | 1091 (52.3) | 5222 (52.2) |
| Partner 5 or more years older | 105 (5.1) | 516 (5.2) |
| Ever pregnant or had a child | 150 (7.3) | 743 (7.4) |
| Prevalent HSV-2 infection | 73 (3.5) | 371 (3.7) |
| Any alcohol use | 173 (8.3) | 817 (8.2) |
| Double or single orphan | 1314 (30.2) | 2638 (27.7) |
| Children’s depression inventory score >=7 | 369 (17.7) | 1795 (17.9) |
| Revised children’s manifest anxiety score >=7 | 570 (27.3) | 2747 (27.5) |
| Partner number |  |  |
| 0 | 1626 (78.9) | 7719 (78.3) |
| 1 | 344 (16.7) | 1687 (17.1) |
| >=2 | 92 (4.5) | 454 (4.6) |

**\*** Missing data in observed: age 0; SES 3; age difference 5, pregnant 23; HSV-2 status 2; alcohol 3; orphan 98; depression 0; anxiety 0; partner number N=24; simulated age difference 5; simulated data missing: partner number 140; alcohol 19; HSV-2 19; pregnant 19; age difference 19; SES 19, orphan 462;

**Appendix table 2:** Controlled direct effect (CDE) of school attendance on incident HIV and HSV-2 by different levels of the mediators partner number and partner age difference setting all HIV cases that had 0 partners to have 1 partner

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **HIV** | | | **HSV-2** | | |
|  | **Risk (%)** | **RD % (95% CI)** | **RR (95% CI)** | **Risk (%)** | **RD %(95% CI)** | **RR (95% CI)** |
| Total effect | | |  |  |  |  |
| High attendance | 5.1 | -2.3 (-2.9, -1.8) | 0.69 (0.75, 0.62) | 7.6 | -9.3 (-10.2, -8.5\_ | 0.45 (0.42, 0.49) |
| Low attendance | 7.4 | 0 | 1 | 17.0 | 0 | 1 |
| CDE No older partners | | |  |  |  |  |
| High attendance | 5.0 | -2.1 (-2.8, -1.4) | 0.71 (0.79 0.63) | 7.6 | -9.3 (-10.0, -8.5) | 0.45 (0.42, 0.49) |
| Low attendance | 7.0 | 0 | 1 | 16.9 | 0 | 1 |
| CDE No sexual partners | | |  |  |  |  |
| High attendance | 0.0 | 0.0 (0.0, 0.0) | 1 (1,1) | 0.0 | 0.0 (0.0, 0.0) | 1 (1,1) |
| Low attendance | 0.0 | 0 | 1 | 0.0 | 0 | 1 |
| CDE One sexual partner | |  |  |  |  |  |
| High attendance | 22.0 | 1. 8 (0.05,3.0) | 1.09 (1.15, 1.02) | 39.0 | -9.4 (-10.7, -8.1) | 0.76 (0.73, 0.79) |
| Low attendance | 20.3 | 0 | 1 | 29.6 | 0 | 1 |

\*RD: Risk Difference, RR: Risk Ratio, CI: Confidence Interval

**Appendix table 3:** Controlled direct effect (CDE) of school attendance on incident HIV and HSV-2 by different levels of the mediators partner number and partner age difference setting all HIV cases that had 0 partners to have 2 partner

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **HIV** | | | **HSV-2** | | |
|  | **Risk (%)** | **RD % (95% CI)** | **RR (95% CI)** | **Risk (%)** | **RD %(95% CI)** | **RR (95% CI)** |
| Total effect | | |  |  |  |  |
| High attendance | 4.6 | -2.3 (-2.9, -1.7) | 0.66 (0.69,0.75) | 7.5 | -9.9 (-10.9, -8.9) | 0.43 (0.39,0.47) |
| Low attendance | 6.9 | 0 | 1 | 17.4 | 0 | 1 |
| CDE No older partners | | |  |  |  |  |
| High attendance | 4.7 | -2.4 (-3.0, -1.7) | 0.66 (0.59, 0.75) | 7.5 | -10.0 (-11.0, -8.9) | 0.43 (0.47, 0.39) |
| Low attendance | 7.0 | 0 | 1 | 17.5 | 0 | 1 |
| CDE No sexual partners | | |  |  |  |  |
| High attendance | 0.0 | 0.0 (0.0, 0.0) | 1 (1,1) | 0.0 | 0.0 (0.0, 0.0) | 1 (1,1) |
| Low attendance | 0.0 | 0 | 1 | 0.0 | 0 | 1 |
| CDE One sexual partner | |  |  |  |  |  |
| High attendance | 8.6 | 1.4 (0.05, 2.3) | 1.12 (1.05, 1.28) | 11.4 | -5.9 (-5.0, -6.8) | 0.66 (0.62, 0.70) |
| Low attendance | 10.1 | 0 | 1 | 17.3 | 0 | 1 |

\*RD: Risk Difference, RR: Risk Ratio, CI: Confidence Interval

**Appendix table 4:** Controlled direct effect (CDE) of school attendance on incident HIV and HSV-2 by different levels of the mediators partner number and partner age difference setting all HIV cases that had 0 partners to have 1 partner and an older partner

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **HIV** | | | **HSV-2** | | |
|  | **Risk (%)** | **RD % (95% CI)** | **RR (95% CI)** | **Risk (%)** | **RD %(95% CI)** | **RR (95% CI)** |
| Total effect | | |  |  |  |  |
| High attendance | 4.6 | -1.7 (-2.3, -1.0) | 0.74 (0.65, 0.83) | 12.8 | -7.0 (-7.8, -6.2) | 0.45 (0.41, 0.50) |
| Low attendance | 6.3 | 0 | 1 | 5.8 | 0 | 1 |
| CDE No older partners | | |  |  |  |  |
| High attendance | 4.5 | -1.1 (-1.7, -0.5) | 0.80 (0.91,0.71) | 12.3 | -0.60 (-6.9, -5.2) | 0.51 (0.46,0.56) |
| Low attendance | 5.6 | 0 |  | 6.0 | 0 | 1 |
| CDE No sexual partners | | |  |  |  |  |  |  |
| High attendance | 3.1 | -0.08 (-0.3, 0.12) | 0.81 (0.70, 0.93) | 5.1 | -3.2 (-4.0, -2.4) | 0.62 (0.55, 0.69) |
| Low attendance | 3.9 | 0 | 1 | 8.3 | 0 | 1 |
| CDE One sexual partner | |  |  |  |  |  |
| High attendance | 6.9 | -1.0 (-1.7, -0.3) | 0.87 (0.80, 0.95) | 17.2 | -8.9 (-9.8, -7.9) | 0.48 (0.45, 0.52) |
| Low attendance | 7.9 | 0 | 1 | 8.3 | 0 | 1 |

\*RD: Risk Difference, RR: Risk Ratio, CI: Confidence Interval