A method for quantifying the regression to the mean effect applied to bivariate binary outcomes in the presence of limited baseline data

Ethan Wilson

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2016

Committee:
James P. Hughes
Barbra A. Richardson

Program Authorized to Offer Degree:
Biostatistics

©Copyright 2016

Ethan A. Wilson

University of Washington

Abstract

A method for quantifying the regression to the mean effect
applied to bivariate binary outcomes in the presence of limited baseline data

Ethan A. Wilson

Chair of the Supervisory Committee:
Dr. James Hughes
Biostatistics

In studies lacking a control group, a crucial step in estimating the study effect is to tease apart the proportion of the total observed change in key outcomes which are due to study participation, from that which is caused by regression to the mean (RTM). We developed novel methods for quantifying RTM effects which jointly model bivariate binary data, while accommodating situations in which baseline data on a representative sample is available for only one of the binary variables. We conducted simulations testing various aspects of our joint model, including cases when modeling assumptions were not met. Using data from a longitudinal cohort study of women at risk for HIV, we then applied our joint model separately to three pairs of bivariate binary outcome measures. We found that weak correlation resulted in a higher proportion of change attributable to study participation for the variable that was not directly selected for, likely due to a lower RTM effect as a result of less predictable selection pressure. Enhanced estimation of RTM effects in non-randomized studies can be obtained using methods which make use of all available data.

Contents

1 Introduction

Regression to the mean (RTM) is a phenomenon that arises when measurements taken over time vary stochastically, leading to a tendency for measures that are more extreme at one visit to approach, on average, an underlying distributional mean at the next visit. For example, suppose an experiment is designed to estimate the change in probability of obtaining heads when tossing fair coins. Suppose that each of 1,000 coins are tossed during the first trial, but only those observed as heads in the first trial are selected for the second trial. While after the first trial the observed proportion of heads among the selected coins is trivially equal to 1, the observed proportion of heads following the second round is almost surely < 1 (the expected proportion of heads is in fact the mean or $\frac{1}{2}$, in the case of a fair coin), resulting in an estimated decrease in the probability of heads among the *selected* coins. Though simple, this example demonstrates how selection for a population based on an extreme measure of a variable can lead to a biased estimation of a change in the mean level of that variable. This same phenomenon occurs among scientifically meaningful measures which are subject to some degree of natural variation and measurement error, and thus it follows that treatment studies involving selection criteria restricted to extreme measures are susceptible to some degree of RTM. Although a well-documented phenomenon, RTM is easily overlooked in the context of treatment effect studies, and quantifying it is of particular importance in studies that lack a control group.

Methods for quantifying the RTM effect have largely been developed and applied in the context of continuous data in which pre- (X_0) and post- (X_1) treatment measures are assumed to have a bivariate normal distribution, including a treatment parameter γ. According to this model we have: $X_1 - \mu = \rho\gamma(X_0 - \mu) + \varepsilon$, with X_0 independent of ε, and ε is normally distributed with mean 0 and variance $\sigma^2(1-\rho^2)$. In the absence of a treatment effect ($\gamma = 1$), X_0 and X_1 have a common mean μ and variance σ^2, with correlation ρ. Likelihood theory has been developed to estimate the RTM effect according to four

common types of sampling designs (see below). In each case, study eligibility criteria may be based on a threshold level $X_0 > k$ (or $X_0 < k$), and the likelihood function for a sample of size n is of the form:

$$L = G(\mu, \sigma^2) \cdot f(X_0) \cdot f(X_1|X_0)$$

$$= G(\mu, \sigma^2) \cdot (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{\frac{1}{2\sigma^2} \sum_{X_0>k} (X_{0i} - \mu)^2\}$$

$$\cdot (2\pi\sigma^2(1-\rho^2))^{-\frac{n}{2}} \exp\{\frac{1}{2\sigma^2(1-\rho^2)} \sum_{X_0>k} (X_{1i} - \mu - \rho\gamma(X_{0i} - \mu))^2\} \quad (1)$$

where $G(\mu, \sigma^2)$ depends on the data available within the sampling scheme [1]. Parameter estimates and standard errors can be obtained using standard numerical methods.

Through the function $G(\cdot)$ the likelihood in (1) accommodates four major sampling designs: *simple* truncated (pre- and post-treatment values are only known for the subjects who are selected for treatment) *censored* (includes the number of subjects who did not meet eligibility criteria), *selected* (includes measured baseline values for subjects who did not meet the eligibility criteria), and *complete* samples (includes the post-treatment values for subjects who did not meet the eligibility criteria)[1]. With each of these successive designs more information can be incorporated into the likelihood function. By including either baseline or full information for subjects below the treatment threshold level, selected and complete sampling designs provide the highest ability to tease apart the effects due to treatment versus RTM.

Methods for normal data have been extended to settings which allow for the underlying mean and variance of measurements to change over time[2]. Beath and Dobson[3] and John and Jawad[4] provide estimation methods which relax the assumptions of normality of X_0 and X_1 (where the contaminating error ε is still assumed to be normal), however the methods were developed in the context of stationary

(untreated) populations. The RTM effect R_k, is quantified in an untreated population according to:

$$R_k = E(X_0 - X_1 | X_0 > k) = (1-\rho)\sigma^2 \left(\frac{h(k)}{1 - H(k)} \right) = (1-\rho)\sigma^2 u(k)$$

where Beath and Dobson use Edgeworth and saddle point approximations to estimate h, the probability density function for X_0 or X_1, while John and Jawad use kernel-based estimation of either h or u. Finally, Muller, Abramson, and Azari[5] further relax the assumption of normality of the contaminating error ε, and apply non-parametric methods to estimate the RTM effect; however these methods were also developed for continuous endpoints in stationary populations.

To our knowledge, very little literature exists on methods for quantifying RTM in the context of non-continuous data. One exception is a method developed for multivariate count data which assumes conditionally independent Poisson outcomes while incorporating eligibility criteria into a baseline conditional distribution, in order to account for the RTM effect[6]. Hughes et al.[7] applied methods for RTM quantification when the outcome is binary, and the data come from a *selected* sampling procedure (i.e. outcome measures are available for a representative sample at baseline, but follow-up measures are only available for a selected sample, based on eligibility criteria).

In this paper we will extend the methods of Hughes et al.[7] to include estimation of the RTM effect among variables that are correlated with, though not (directly) used in, eligibility screening measures. While for these variables there are no baseline data available for a representative sample, they are *indirectly selected* (e.g. if the selection criteria is "at least one unprotected sex act in the past 6 months" then the proportion of individuals who had "unprotected sex at their most recent act" is likely higher than background). We propose a novel method for estimating the RTM effect in these indirectly selected measures which utilizes available information on the relationship between the directly selected and indirectly selected measures.

2 Statistical Methods

Consider estimation of both treatment and RTM effects for two binary variables X and Y measured over time, with respective joint distributions (as in[7]) represented by:

$$P(X_t, X_{t-1}, ..., X_1, X_0) = P(X_t|X_{t-1}) \cdot P(X_{t-1}|X_{t-2}) \cdot ... \cdot P(X_1|X_0)P(X_0)$$

$$P(Y_t, Y_{t-1}, ..., Y_1, Y_0) = P(Y_t|Y_{t-1}) \cdot P(Y_{t-1}|Y_{t-2}) \cdot ... \cdot P(Y_1|Y_0)P(Y_0)$$

so that the respective probabilities of X and Y depend only on the most recently measured value of that variable. Suppose also that for a given time k we have (throughout we shall adopt the subscript c to indicate conditional arguments):

$$\text{logit}(p_{x,c}) \equiv \text{logit}(P(X_k|X_{k-1}, E_k)) = \alpha_0 + \alpha_1 X_{k-1} + \alpha_2 E_k \tag{2}$$

$$\text{logit}(p_{y,c}) \equiv \text{logit}(P(Y_k|Y_{k-1}, E_k)) = \gamma_0 + \gamma_1 Y_{k-1} + \gamma_2 E_k \tag{3}$$

where E_k is an indicator that the subject was on the study from time k-1 to k. Here α_0 and γ_0 reflect the underlying baseline prevalences, α_1 and γ_1 are autocorrelative parameters, and α_2 and γ_2 are study effect parameters. As is shown in [7], the marginal off- and on-study probabilities of X and Y in the representative population are predicted as (assuming sufficient time has passed for X and Y to reach a steady state after study exposure):

$$p_{x,off} = \frac{e^{\alpha_0}(1 + e^{\alpha_0+\alpha_1})}{1 + 2e^{\alpha_0} + e^{2\alpha_0+\alpha_1}} \qquad p_{y,off} = \frac{e^{\gamma_0}(1 + e^{\gamma_0+\gamma_1})}{1 + 2e^{\gamma_0} + e^{2\gamma_0+\gamma_1}} \tag{4}$$

$$p_{x,on} = \frac{e^{\alpha_0+\alpha_2}(1+e^{\alpha_0+\alpha_1+\alpha_2})}{1+2e^{\alpha_0+\alpha_2}+e^{2\alpha_0+\alpha_1+2\alpha_2}} \qquad p_{y,on} = \frac{e^{\gamma_0+\gamma_2}(1+e^{\gamma_0+\gamma_1+\gamma_2})}{1+2e^{\gamma_0+\gamma_2}+e^{2\gamma_0+\gamma_1+2\gamma_2}} \tag{5}$$

We may quantify the strength of association between an arbitrary X and Y using the odds ratio (OR):

$$\psi \equiv \frac{p_{11} \cdot p_{00}}{p_{01} \cdot p_{10}}$$

where p_{ij} is the probability that $X = i$ and $Y = j$. More generally, conditioning on the previous time we have:

$$\psi_c^{rs} \equiv \frac{P(XY_k = 11|XY_{k-1} = rs) \cdot P(XY_k = 00|XY_{k-1} = rs)}{P(XY_k = 01|XY_{k-1} = rs) \cdot P(XY_k = 10|XY_{k-1} = rs)} \equiv \frac{p_{11|rs} \cdot p_{00|rs}}{p_{01|rs} \cdot p_{10|rs}} \tag{6}$$

where $p_{ij|rs}$ is the probability that$(X, Y) = (i, j)$, conditional on the previously measured $(X, Y) = (r, s)$. We assume that $\psi_c^{rs} = \psi_c$ for all r, s –in other words the strength of association between X and Y, conditional on the previous values, is equivalent for all previously measured (X, Y)– and that ψ_c is constant over time and independent of both variable treatment effects (In sensitivity analyses, we will relax the first assumption by including four ψ_c^{rs} parameters for each conditional state.)

Now we consider modeling the joint distribution of (X, Y) for t+1 discrete times:

$$P(XY_t, XY_{t-1}, ..., XY_1, XY_0) = P(XY_t|XY_{t-1}) \cdot P(XY_{t-1}|XY_{t-2}) \cdot ... \cdot P(XY_1|XY_0)P(XY_0) \tag{7}$$

Each $P(XY_k|XY_{k-1})$ on the right hand side of (7) can then be expressed as a function of the marginal conditional distributions of X (2) and Y (3) and the conditional OR (6). For example, let $P(XY_k = ij|XY_{k-1} = rs) \equiv p_{ij|rs}$ (for all k), and note that $p_{i\cdot|rs} \equiv P(X_k = i|XY_{k-1} = rs)$ and $p_{\cdot j|rs} \equiv P(Y_k = j|XY_{k-1} = rs)$. Assuming $P(X_k = i|XY_{k-1} = rs) = P(X_k = i|X_{k-1} = r)$ and $P(Y_k = j|XY_{k-1} = rs) =$

$P(Y_k = j|Y_{k-1} = s)$ – i.e. the probability of X (or Y) is independent of the previously measured Y (or X), conditional on the previously measured X (or Y) – then for $p_{11|rs}$ we have (noting that (8) is implied by the range restriction of $p_{11|rs}$):

$$p_{11|rs} = \frac{p_{01|rs} \cdot p_{10|rs} \cdot \psi_c}{p_{00|rs}} = \frac{(p_{\cdot 1|rs} - p_{11|rs}) \cdot (p_{1\cdot|rs} - p_{11|rs}) \cdot \psi_c}{(1 - p_{\cdot 1|rs} - p_{1\cdot|rs} + p_{11|rs})}$$

$$= \frac{(p_{\cdot 1|s} - p_{11|rs}) \cdot (p_{1\cdot|r} - p_{11|rs}) \cdot \psi_c}{(1 - p_{\cdot 1|s} - p_{1\cdot|r} + p_{11|rs})}$$

$$= \frac{(p_{y,c} - p_{11|rs}) \cdot (p_{x,c} - p_{11|rs}) \cdot \psi_c}{(1 - p_{y,c} - p_{x,c} + p_{11|rs})}$$

$$\Leftrightarrow p_{11|rs}^2(\psi_c - 1) - p_{11|rs}\{1 + (p_{y,c} + p_{x,c})(\psi_c - 1)\} + \psi_c p_{y,c} p_{x,c} = 0$$

$$\Leftrightarrow p_{11|rs} = \frac{\{1 + (p_{y,c} + p_{x,c})(\psi_c - 1)\} \pm \sqrt{\{1 + (p_{y,c} + p_{x,c})(\psi_c - 1)\}^2 - 4(\psi_c - 1)\psi_c p_{y,c} p_{x,c}}}{2(\psi_c - 1)}$$

$$\Rightarrow p_{11|rs} = \frac{\{1 + (p_{y,c} + p_{x,c})(\psi_c - 1)\} - \sqrt{\{1 + (p_{y,c} + p_{x,c})(\psi_c - 1)\}^2 - 4(\psi_c - 1)\psi_c p_{y,c} p_{x,c}}}{2(\psi_c - 1)} \quad (8)$$

Similar algebra may be used to calculate the remaining conditional probabilities: $p_{10|rs}$, $p_{01|rs}$, and $p_{00|rs}$. Finally, in order to complete the full probabilistic model in (7) we need only to derive the four marginal baseline probabilities: $P(XY_0 = ij) \equiv p_{ij}$ (see appendix for derivation). Using our formulations for $p_{ij|rs}$ and p_{ij} we can obtain parameter estimates for $(\alpha_0, \alpha_1, \alpha_2, \gamma_0, \gamma_1, \gamma_2, \tau)$ ($\tau \equiv \log(\psi_c)$), by maximizing the log-likelihood:

$$\ell = \sum_{i=0}^{1}\sum_{j=0}^{1} n_{ij,0} \cdot \log(p_{ij}) + \sum_{k=1}^{t}\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{r=0}^{1}\sum_{s=0}^{1} n_{ij|rs,k} \cdot \log(p_{ij|rs})$$

where any missing data is handled by summing over all of the missing X and Y values in the joint probability. Using the estimated parameters, we are able to estimate the proportion of the observed

change in prevalence of X or Y that is attributable to study participation as:

$$\delta \equiv \frac{p_{off} - p_{on}}{p_{BL} - p_{on}} \tag{9}$$

where p_{off} and p_{on} can be estimated according to (4) and (5) respectively, and p_{BL} is the observed prevalence at baseline. Standard errors for derived parameters $(\delta, p_{off}, p_{on})$ are obtained using standard deviations from 200 bootstrap simulations.

2.1 Simulation study

In order to test various features of our proposed model we ran a simulation study in which we simulated combinations of the following derived parameters:

- $p_{x,off}$ & $p_{y,off} = \{0.6, 0.8, 0.9\}$
- ϕ_x & $\phi_y = e^{\alpha_1}$ & $e^{\gamma_1} = \{2, 6\}$
- θ_x & $\theta_y = e^{\alpha_2}$ & $e^{\gamma_2} = \{1, \frac{5}{6}, \frac{2}{3}\}$
- $\psi_c = e^{\tau} = \{1, 1.5, 10.0\}$

Our motivation for these parameter values was driven by their scientific relevance, while providing insight to the statistical properties of our model. Including all combinations of these parameter values yielded a total of 972 distinct settings, each of which included 1000 simulations (see appendix for simulation methods).

While the mechanisms by which a study may influence the marginal probabilities of two binary variables are readily hypothesized, the extent to which a study may affect the strength of association between two binary variables is less clear. Our joint model relies on the assumption that ψ_c is independent of study

participation (i.e. constant over time) however, it may be hypothesized that this is not true. For this reason, we investigated how the statistical properties of our model are affected when this key assumption is violated. We conducted additional simulations in which data were generated using two distinct conditional ORs, depending on whether or not a subject was participating in the study in the preceding 6 months. In these simulations, $\psi_{c,off}$ (conditional OR if a subject was not participating in the study in the preceding 6 months) and $\psi_{c,on}$ (conditional OR if a subject was participating in the study in the preceding 6 months) could take on the following range of values: {1.05, 1.10, 1.20, 1.30, 1.50, 1.75, 2.00, 2.50, 5.00, 10.0}. We tested all combinations of these values for both $\psi_{c,-}$ parameters (omitting the cases when the two ORs are equivalent) and tested the range of treatment parameter values $\theta_y = \{1, \frac{5}{6}, \frac{2}{3}\}$, while fixing the remaining parameter values accordingly: $p_{x,off} = 0.75$, $\phi_x = 6.63$, $\theta_x = 0.96$ $p_{y,off} = 0.72$, $\phi_y = 2.83$.

2.2 HPTN 064 data

Hughes et al.[7] developed and applied methods to quantify the amount of change in a risk behavior – unprotected sex in the past 6 months – that is associated with study participation versus RTM, using data obtained from the HIV Prevention Trials Network Women's HIV SeroIncidence Study HPTN 064[8]. In this study outcomes of interest were obtained via participant completion of audio computer-assisted self-interviews (ACASI) at baseline and each follow-up visit.

Define:

$$X = \begin{cases} 1 & \text{if a participant reports unprotected sex (with a man, in the past 6 months)} \\ 0 & \text{otherwise} \end{cases}$$

for X measured at study baseline, and 6 and 12 months of follow-up. Full or partial X data were available

for subjects in the *selected* population, that is the population of subjects who were enrolled, and therefore eligible to participate in the study. Because unprotected sex was part of the study eligibility criteria, an estimate of the baseline prevalence of X was available in the *unselected* population, i.e. the population of subjects who were screened and eligible to participate in the study based on all criteria *except* that defined by X.

Here, we expand on the univariate analysis of X by modeling (X, Y) jointly, where Y may be any other binary outcome of interest. While quantifying the RTM effect in the univariate model relies on an unbiased estimation of variable prevalence in the *unselected* population, our joint model can be applied even when data for Y are only available among the *selected*, or enrolled population, provided that full data are available for X, as described above. We created separate joint (X, Y) models for each of three binary variables using participants self-reported answers to the following questions:

- Y_1: In the past 6 months, the last time you had vaginal sex with a man did you use a condom?
- Y_2: In the past 6 months, have you been concerned about having enough food for you and/or your family?
- Y_3: In the past 6 months, did you ever need medical care but could not get it?

We specifically included these variables so that we could apply our method to bivariate relationships that were hypothesized to have higher (Y_1) and lower (Y_2 and Y_3) positive correlation with X.

It is easy to see that, theoretically, Y_1 is nested within X, and that $P(Y_1 = 1) = P(Y_1 = 1|X = 1) \cdot P(X = 1)$. This relationship implies that an unbiased estimate of $p_{y,off}$ can be obtained using the observed baseline prevalence of $(Y_1 = 1|X = 1)$, and the estimate of $p_{x,off}$ from the *unselected* population (assuming $P(Y_1 = 1|X = 1)$ is independent of study enrollment). To illustrate this method we include

one additional model for the (X, Y_1) data, which accounts for this deterministic relationship. Our second model is similar to that of (7), with the exception that:

$$P(XY_k|XY_{k-1}) = P(Y_k|X_k, XY_{k-1}) \cdot P(X_k|XY_{k-1}) \tag{10}$$

$$= \begin{cases} p_{y_k|x_k,h} \cdot p_{x_k|h} & (X,Y)_k = (1,1) \\ (1 - p_{y_k|x_k,h}) \cdot p_{x_k|h} & (X,Y)_k = (1,0) \\ (1 - p_{kx|h}) & (X,Y)_k = (0,0) \end{cases}$$

where h denotes $(X, Y)_{k-1}$ and

$$\text{logit}(p_{x_k|h}) = \alpha_0 + \alpha_{1j} I_{[h=(1,j)]} + \alpha_2 E_k$$

$$\text{logit}(p_{y_k|x_k,h}) = \gamma_0 + \gamma_{1j} I_{[h=(1,j)]} + \gamma_2 E_k.$$

This is an 8 parameter model in which the α_{1j} parameters are based on the previously measured (X, Y) only, while the γ_{1j} parameters additionally account for the current X measure. This model assumes that $P(X = 0, Y = 1) = 0$, and while there were some subjects in the HPTN 064 data set for which $(X, Y_1) = (0, 1)$, overall prevalence was low (1.9%). To account for these instances, we treated the (0,1) values as missing, and all missing data were handled by summing over the joint probability in the log-likelihood.

3 Results

3.1 Simulation study

The results of our primary simulation study including 972 total settings are presented in table 1 and the appendix (tables A1-A3). For the Y variable treatment parameter, γ_2, we estimated the bias, coverage probability, power, type I error, and standard error; while also exploring any trends in these properties as a function of the parameters. We observed no apparent trends in the bias of γ_2 (tables A1-A3), which quantifies the effect of study participation on the Y variable (mean relative bias = -0.003, range = [-0.051, 0.033]). The coverage probability for γ_2 was also consistent across all parameters (mean coverage probability = 0.970, range = [0.946, 0.985]). In the case of no treatment effect ($e^{\gamma_2} = 1$), there were no observed trends in bias (absolute mean bias < 0.001, range = [-0.010, 0.005]), and the type I error rate (α) was at or below nominal levels (mean $\alpha = 0.029$, range = [0.016, 0.050]).

All power estimates refer to testing H_0 ($\gamma_2 = 0$), and are presented in table 1. Because we observed no trends in mean power for both the X autocorrelative parameter (mean power for γ_2 averaged over simulations with $e^{\alpha_1} = (2, 6)$ were 0.872 and 0.875, respectively) and the X treatment parameter (mean power for γ_2 averaged over simulations with $e^{\alpha_2} = (\frac{2}{3}, \frac{5}{6}, 1)$ were 0.874, 0.874, and 0.873, respectively) the power estimates presented in the table are averaged over these parameters. When the treatment effect size is lower ($e^{\gamma_2} = \frac{5}{6}$) the power for γ_2 increases with the autocorrelative parameter γ_1 and the off-study probability of X ($p_{x,off}$), while it decreases with $p_{y,off}$ and ψ_c. When the treatment effect size is high ($e^{\gamma_2} = \frac{2}{3}$) the power remains close to 1 across nearly all parameters, with the exception of the case when $p_{x,off} = 0.6$, $p_{y,off} = 0.9$, and $\psi_c = 10$ (mean power for $\gamma_2 = 0.82$).

Standard error estimates for γ_2 were consistent across the three values simulated, and they tended to increase with $p_{y,off}$ and ψ_c and decrease with $p_{x,off}$ and γ_1, which was also consistent with the power results

obtained when the treatment effect was low ($e^{\gamma_2} = \frac{5}{6}$). The relationship between $p_{x,off}$ with both standard error and power estimates for γ_2 is likely explained by the larger on study sample size of Y that was a result of the higher background prevalence of the selection variable, X.

While our simulations did not produce standard error estimates for the respective pre- and post-study prevalence $p_{y,off}$ and $p_{y,on}$, our simulations yielded very low biases for these parameters, with an absolute mean bias <0.001 in both $p_{y,off}$ (range= [-0.002, 0.002]) and $p_{y,on}$ (range = [-0.003, 0.003]).

3.1.1 Violation of assumption that ψ_c is constant over time

Bias estimates for γ_2 from our simulations in which the assumption of a constant conditional odds ratio was violated are provided in table 2. Bias estimates were consistent in both sign and magnitude across each level of the treatment parameter ($e^{\gamma_2} = \frac{2}{3}, \frac{5}{6}$, and 1). For this reason, the estimates presented in table 2 are averaged over the three treatment levels simulated. As expected, the magnitude of the bias and type I error increases as a function of $|\psi_{c,off} - \psi_{c,on}|$. We observe an overestimate of γ_2 when $\psi_{c,off} > \psi_{c,on}$ and an underestimate when $\psi_{c,off} < \psi_{c,on}$. These results demonstrate that large violations of the assumption of constancy of ψ_c can result in substantial bias.

3.2 HPTN 064 study

Full baseline information for X (any unprotected sex in the past 6 months) was available for a representative sample of 4126[1] women, of which 2099 (50.9%) participants in the study comprised the *selected* sample based on X. Of the 2099 enrolled participants, follow-up information was available for 1953 (93.0%) and 1525 (72.3%) subjects after 6 and 12 months of follow-up, respectively. Empirical and model-based prevalence estimates and missing data information are presented in table 3. Among study participants, the

[1] adjusted sample size described in [7]

absolute decline in prevalence from baseline to 12 months of follow-up was 0.19 for any reported unprotected sex, 0.19 for reported unprotected sex at most recent act, 0.10 for reported food concerns, and 0.06 for reported foregone medical care; thus the effect of study participation appeared beneficial for each of these variables.

Table 4 presents the results of our method applied to each of the 3 bivariate models. Our estimates of the observed change that is attributable to study participation were 69% for reported last sex as unprotected (95% bootstrap C.I. = (63%, 74%)), 95% for food insecurity (95% bootstrap C.I. = (87%, 103%)), and 98% for foregone medical care (95% bootstrap C.I. = (87%, 109%)); suggesting that there is a beneficial study effect for each of these variables, while accounting for any change in prevalence due to RTM. While estimates for $p_{x,off}$ were consistent across all models (0.750), the estimate of $p_{x,on}$ (and therefore δ_x) was lower for the model that included unprotected sex at most recent act. Despite this discrepancy all three analyses indicate a minor effect of treatment on the prevalence of any unprotected sex; in each model the proportion of this change in prevalence which is attributable to study participation is not significantly different from 0, at the $\alpha = 0.05$ level. Using the alternative model derived in (10), the estimated off-study (*unselected*) probability of unprotected sex at the most recent act was 0.630, substantially lower then the corresponding estimate from our primary model of 0.736. Using this model we estimated 46% of the change in prevalence was attributable to study participation (95% bootstrap C.I. = (41%, 51%)).

Over the course of the study we expect the prevalence of the binary variables to approach a steady state, which we quantify as p_{on}, at which point the combined effects of RTM and study participation will have reached an equilibrium. Assuming sufficient time has elapsed to reach this state, we expect our model-derived prevalence estimate ($\hat{p}_{on}$) to be approximately equal to the empirically observed 12 month prevalence. While this was the case for the foregone medical care and food insecurity variables (difference between empirical and model-derived prevalence = 0.008 and 0.016, respectively), this difference was much

greater (0.630 observed, versus 0.554 model-predicted) for the last unprotected sex act variable. We notice as well that this variable has a significantly higher degree of missing data that was attributable to failure to respond on the ACASI questionnaire, as opposed to being lost to follow-up (table 3). For this reason, we explored two imputation methods for the prevalence of unprotected sex at the most recent act at 6 and 12 months of follow-up. The first method uses the empirically observed (non-missing) prevalence of "unprotected sex at most recent act" given "any unprotected sex", and handles each follow-up time separately. The second method imputes all missing data using the model-derived estimates of the prevalence of each possible bivariate state across all three times. We observe that the two imputation methods handle the missing data nearly identically, resulting in 12 month prevalence estimates of 57.1% using empirical imputation and 57.8% using the model-based imputation. These results are concordant with our expectation that the 12 month prevalence (after data imputation) is closer to our model estimate of $p_{y,on}$ (55.4%).

Upon closer investigation, we observed that the percent of times that subjects reported no unprotected sex, given that the response to the unprotected sex at the most recent act was missing, were 91% and 94% at 6 and 12 months, respectively. The preponderance of missing information for the ACASI question about unprotected sex at the most recent act (compared to the lower rates of missing data among other variables measured) and the high proportion of times that this missing information was concurrent with self-report of no unprotected sex in the past 6 months, suggests that subjects may have failed to respond to the "unprotected sex at most recent act" question, simply because it was already implied by their response to the "any unprotected sex" question. Assuming there are no alternative mechanisms underlying the missingness in our data, our method for handling missing data is able to account for this observed pattern.

4 Discussion

Regression to the mean has been characterized in the context of four major sampling schemes: *simple*, *censored*, *selected*, and *complete*. In the case of binary data, there is in fact no distinction between the censored and selected sampling methods, since knowledge about the number of subjects not meeting a binary variable requirement is equivalent to full information regarding that variable. In this paper we developed novel methods for quantifying RTM in the context of what we call *indirectly selected* sampling, i.e. sampling for a variable Y which arises as a result of *selected* sampling of X, due to the correlation between X and Y. We applied our methods to the HPTN 064 study data, finding that among binary measures subject to indirect selection, a majority of the observed decline in prevalence was attributable primarily to study participation. Not surprisingly "unprotected sex at the most recent act" was highly correlated with "any unprotected sex", resulting in a non-trivial RTM effect and thus a lower proportion of the observed change in prevalence was attributable to study participation. The two remaining variables were largely uncorrelated with the selection variable, implying that the baseline and off-study prevalence were nearly identical and the RTM effect was low. Interestingly, while nearly no study effect was observed in the "any unprotected sex" variable, we observed a substantial study effect for the "unprotected sex at the most recent act". Though speculative, these results could suggest that participant awareness of the latter ACASI question resulted in an increased use of condoms in the immediate acts prior to each follow-up visit, but seldom resulted in comprehensive condom use.

We also proposed a variant of our bivariate binary model which utilizes the conditional probability of the indirectly selected variable, given the current measure of the directly selected variable (10). A benefit of using this conditional model in the case of *nested* binary variables, is that an estimate of the off-study prevalence of the indirectly selected variable can be obtained with fewer assumptions than the model in (7). When we applied this conditional model to the variable "unprotected sex at the most recent act"

we noticed discrepancies in the derived parameter estimates, compared to our original model. This is not entirely surprising, given that the reported data had to be altered in order to reflect nesting; as a result we can only speculate on the accuracy of these estimates.

It is important to note that, while our primary model is more general, the extent to which model misspecification can result in a lack of precision or even biased estimation is unclear. For example, when binary variables are nested in truth (as was the case for two of HPTN 064 variables), the underlying odds ratio is infinite and thus may be an obsolete and potentially incorrect measure of the association between two variables. The conditional model presents an alternative which accounts for the relationship between the selected measures without using the odds ratio. While we did not do so here, this conditional model may also be generalizable to the case of non-nested variables, however it would require additional parameters which we found less interpretable.

A limitation of our model is that, without additional knowledge regarding the underlying off-study prevalence of the indirectly selected Y measures, we cannot definitively test whether two key assumptions are violated. We assume the probability of enrollment (conditional on study eligibility) is independent of each Y outcome, and that the conditional odds ratio between X and Y is independent of study participation. Our second simulation study was able to provide a gauge for the biases that can arise when ψ_c is not independent of study participation. Still, for this study it seems reasonable to assume that the set of Y variables were not related to study enrollment, conditional on X. We were able to relax the assumption that ψ_c is equivalent for each previous (X, Y) state by including 4 total ψ_c^{rs} parameters for each conditional state in (6). However, this resulted in trivial attenuations of our estimate of δ_y (eq. 10) for last sex unprotected (-1.61), food insecurity (-3.66), and foregone medical care (+2.74); at the cost of three additional nuisance parameters. Simulating data according to these extra parameters was also beyond the scope of our investigation, since it would involve a far more complex set of simulations; further choice of parameter

values for each of the four conditions was less intuitive.

An alternative model design that we considered was one which allows for overdispersion in the underlying probability of the binary variables, based on the premise that a subject's tendency to engage in risky behavior may not be constant over time. This type of mechanism would lead to a RTM phenomenon that is more analogous to that which is observed in continuous data. A key distinction of this model is that the true underlying off-study prevalence in the study population is not necessarily equivalent to that of *unselected* (representative) population, as is the case in our model that assumes no extra-binomial variation. If in truth there is overdispersion, than our estimation of p_{off} in the *selected* population is biased, which (assuming selection occurs such that study baseline prevalence is higher than that of the *unselected* population) would lead to an underestimate of δ in (9). However, upon exploration of a beta-binomial model we found that without at least one additional baseline measurement the RTM effect becomes analytically intractable – the model lacks sufficient information to distinguish whether or not observed changes in prevalence are a result of beta-binomial variation or treatment. Still, provided sufficient baseline data exists, development of RTM quantification methods which incorporate extra-binomial variation may be an important area of future research, especially when applied to situations for which there is suspected binomial overdispersion.

Data collection costs and observational design are two of many reasons why studies often lack sufficient information required to characterize RTM effects in target outcomes. This lack of data, coupled with the ubiquity of RTM, creates a need for innovative methods, which fully utilize the information available. By drawing on their relationship with variables subject to *selected* sampling, our method serves to reframe the sampling mechanism for data that would otherwise be classified as arising from a *simple* sampling design. We are aware of no previous method that quantifies RTM in this context of *indirecty selected* sampling. In the future, our method for quantifying RTM in the case of indirectly selected sampling could also be

extended to the case where the variables that are selected, indirectly selected, or both, are either continuous or count data.

5 Appendix

5.1 Derivation of baseline P(XY_0)

Noting that the number of observations in all possible (X, Y) states follows a multinomial distribution with four possible categories, we can derive the steady state (marginal) probabilities of all four (X, Y) outcomes in the representative population, using the following proof (here we adopt the notation $XY_{k,ij}$ as an indicator of $(X, Y) = (i, j)$ at time k):

$$E(XY_{0,ij}) = E(E(XY_{0,ij}|XY_{-1,rs}))$$

$$\Leftrightarrow [p_{00}\ p_{01}\ p_{10}\ p_{11}] =$$

$$\sum_{r=0}^{1}\sum_{s=0}^{1} E\{[\ E(XY_{0,00}|XY_{-1,rs}) \quad E(XY_{0,01}|XY_{-1,rs}) \quad E(XY_{0,10}|XY_{-1,rs}) \quad E(XY_{0,11}|XY_{-1,rs})\ ]\} \qquad \text{(A1)}$$

where for each element of the vector on the right hand side (RHS), we have:

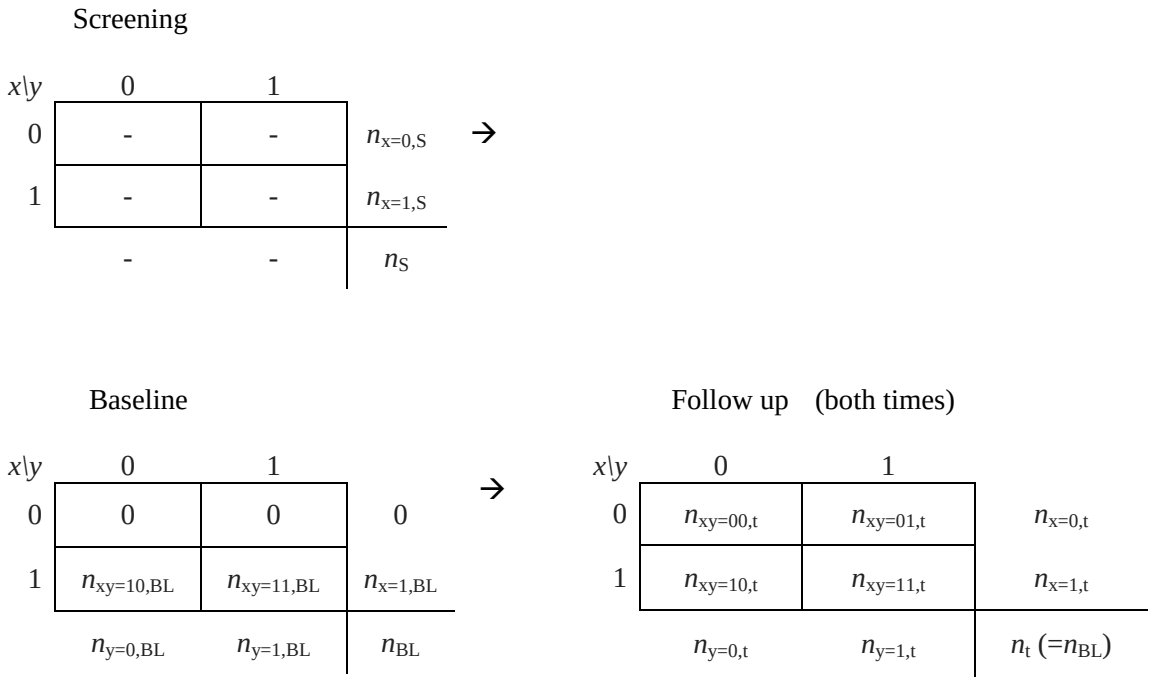
$$\sum_{r=0}^{1}\sum_{s=0}^{1} E\{[\ E(XY_{0,ij}|XY_{-1,rs})\} = p_{ij|00}\cdot p_{00} + p_{ij|01}\cdot p_{01} + p_{ij|10}\cdot p_{10} + p_{ij|11}\cdot p_{11} = p_{ij}.$$

By substituting (1-p_{01}-p_{10}-p_{11}) for p_{00} in (A1) and omitting the respective first elements of the LHS and RHS, we obtain a system of three linear equations that can be solved to obtain the unconditional probabilities: p_{01}, p_{10}, and p_{11} (and thus also p_{00}).

5.2 Simulation Methods

We will generate bivariate binary data (X_i, Y_i) at screening (i=0.1), baseline (i=0.2), and two follow-up times (i= 1, 2) according to the model derived in the Statistical Methods section. In each simulation we assume that full information is known about X at all four of these times; while information about

Y is assumed unknown at screening, but known at baseline and the two follow-up times. Those who are eligible at screening are all people such that $X=1$, regardless of Y. The following flow chart represents what is known for both grouped data (X and Y together) and marginal data (only X or Y):



Our goal is to simulate data according to prespecified values of the following parameters (described in the Statistical Methods section): α_0, α_1, & α_2 (2); γ_0, γ_1, & γ_2 (3); and, ψ_c (6). The first step is to equate these parameters with meaningful values of $p_{x,off}$ & $p_{y,off}$ (4) (the respective off-study probabilities for X and Y), ϕ_x and ϕ_y (a measure of the autocorrelation over consecutive measurements for X and Y) and θ_x and θ_y (for X and Y respectively, the OR comparing two groups with the same previous X (or Y) measure, differing only in: study participation during the time since the previous visit). Our simulations included

all combinations for each of the following values:

- $p_{x,off}$ & $p_{y,off} = \{0.6, 0.8, 0.9\}$
- ϕ_x & $\phi_y = \{2, 6\}$
- θ_x & $\theta_y = \{1, \frac{5}{6}, \frac{2}{3}\}$
- $\psi_c = \{1.0, 1.5, 10\}$

for which parameter values are derived using the relationships described below.

5.2.1 Obtaining Parameters

α_1 & γ_1 from ϕ_x & ϕ_y:

We start by noting that the value $\phi_x \equiv \frac{p_x00 \cdot p_x11}{p_x01 \cdot p_x10} = e^{\alpha_1}$ (where $p_xij \equiv P(X_k, X_{k+1}) = (i, j)$ for consecutive times k and k+1). Proof: from [7] we have (for the unselected or off study population): $P(X_0, X_{-1}) = P(X_0|X_{-1})P(X_{-1})$

$$= \begin{cases} \frac{e^{2\alpha_0+\alpha_1}}{1+2e^{\alpha_0}+e^{2\alpha_0+\alpha_1}} & (X_0, X_{-1}) = (1, 1) \\ \frac{e^{\alpha_0}}{1+2e^{\alpha_0}+e^{2\alpha_0+\alpha_1}} & (X_0, X_{-1}) = (0, 1) \text{ or } (1, 0) \\ \frac{1}{1+2e^{\alpha_0}+e^{2\alpha_0+\alpha_1}} & (X_0, X_{-1}) = (0, 0) \end{cases}$$

And $\frac{p_x00 \cdot p_x11}{p_x01 \cdot p_x10}$ simplifies to $\frac{e^{2\alpha_0+\alpha_1}}{e^{2\alpha_0}} = e^{\alpha_1}$. (Similarly for Y we have $\phi_y = e^{\gamma_1}$).

α_0 & γ_0 from $(p_{x,off}, \alpha_1)$ & $(p_{y,off}, \gamma_1)$:

Given initial auto-correlative parameters (α_1 and γ_1) as well as initial marginal probabilities $p_{x,off}$ and $p_{y,off}$ and using the relationship in (4), we can derive either α_0 or γ_0 as the log of the positive root of a quadratic equation.

α_2 & γ_2 from θ_x & θ_y:

Equations (2-3) in the Statistical Methods section show that $\theta_x \equiv e^{\alpha_2}$ (and $\theta_y \equiv e^{\gamma_2}$) are interpreted as the OR comparing two groups with the same previous X (or Y) measure, differing only in study participation during the time since the previous visit. Thus α_2 and γ_2 were chosen based on ORs for measuring either variable treatment effect.

5.2.2 Simulate Data

Starting with a sample of $n = 4000$ subjects, we created screening data by randomly simulating Multinomial$(n, \vec{p} = [p_{00}, p_{01}, p_{10}, p_{11}])$ data with $p_{ij} = P(X, Y) = (i, j)$ in the *unselected* population (quantification of each p_{ij} requires the parameters $\alpha_0,\ \alpha_1,\ \alpha_2,\ \gamma_0,\ \gamma_1,\ \gamma_2,\ \psi_c$ and is derived in the Statistical Methods section). Using only data corresponding to $X = 1$ ($n_{10,S}$ and $n_{11,S}$), we reduced each category by a factor of $p_{enroll} = \frac{2099}{3234}$ (among women who were eligible for the study in [7] the observed proportion who enrolled), yielding baseline data ($n_{10,BL}$ and $n_{11,BL}$). We then simulated data at the first follow-up time. Due to selection criteria $n_{00,BL}$ and $n_{01,BL}$ were both 0, and we thus simulated two sets of multinomial data, each including probabilities that were conditional on both study participation as well as previous multinomial category, according to (note that $\vec{n}_{t1} = \vec{n}_{t1|10} + \vec{n}_{t1|11}$):

$$\vec{n}_{t1|10} = [n_{00,t1}\ n_{01,t1}\ n_{10,t1}\ n_{11,t1}]|n_{10,BL} \sim M(n_{10,BL}, \vec{p}_{|10})^{\dagger}$$

$$\vec{n}_{t1|11} = [n_{00,t1}\ n_{01,t1}\ n_{10,t1}\ n_{11,t1}]|n_{10,BL} \sim M(n_{11,BL}, \vec{p}_{|11})^{*}$$

Last we simulate data for the second follow-up time, where we now have 4 non-zero categories from follow-up time 1, yielding:

$$\vec{n}_{t2|00} = (n_{00,t2}, n_{01,t2}, n_{10,t2}, n_{11,t2})|n_{00,t1} \sim M(n_{00,t1}, \vec{p}_{|00})^{*}$$

$$\vec{n}_{t2|01} = (n_{00,t2}, n_{01,t2}, n_{10,t2}, n_{11,t2})|n_{01,t1} \sim M(n_{01,t1}, \vec{p}_{|01})^{*}$$

$$\vec{n}_{t2|10} = (n_{00,t2}, n_{01,t2}, n_{10,t2}, n_{11,t2})|n_{10,t1} \sim M(n_{10,t1}, \vec{p}_{|10})$$

$$\vec{n}_{t2|11} = (n_{00,t2}, n_{01,t2}, n_{10,t2}, n_{11,t2})|n_{11,t1} \sim M(n_{11,t1}, \vec{p}_{|11})$$

5.2.3 Input Data

The input data for the likelihood reflects all known quantities and, for concision, we grouped redundant values (e.g., $n_{t1|10}$ and $n_{t1|11}$ can be respectively added to $n_{t2|10}$ and $n_{t2|11}$ based on our assumption that $P(X, Y)$ depends only on the most recently measured value of (X, Y), when available). We thus yield 20 distinct values that can be incorporated into the likelihood:

$$input\ data = \{n_{x=0,S},\ n_{x=1,S},\ n_{10,BL},\ n_{11,BL},\ \vec{n}_{t2|00},\ \vec{n}_{t2|01},\ \vec{n}_{t1+t2|10},\ \vec{n}_{t1+t2|11}\}$$

[†]Derivations for each of the 4 elements of the 4 vectors of conditional probabilities ($\vec{p}_{|00}$, $\vec{p}_{|01}$, $\vec{p}_{|10}$, $\vec{p}_{|11}$) can be found in the statistical methods section

5.2.4 Likelihood optimization

The likelihood is described in detail in the Statistical Methods section. All conditional and marginal probabilities for both X and Y, in addition to conditional and marginal probabilities for multinomial (X, Y) pairs can be expressed as function of our 7 parameters, exclusively (in each case conditioning means on the value of that variable at the previous time). With the input data available described in 3.1.3 we incorporate the following probabilities, corresponding to the data:

	Input data	Corresponding p (or $\vec{p}$)
	$n_{x=0,S}$	$1 - p_{x,off}$
Off	$n_{x=1,S}$	$p_{x,off}$
Study	$n_{10,BL}$	$\frac{p_{10}}{p_{x,off}}$
	$n_{11,BL}$	$\frac{p_{11}}{p_{x,off}}$
	$\vec{n}_{t1,t2\|00}$	$\vec{p}_{\|00}$
On	$\vec{n}_{t1,t2\|01}$	$\vec{p}_{\|01}$
Study	$\vec{n}_{t2\|10}$	$\vec{p}_{\|10}$
	$\vec{n}_{t2\|11}$	$\vec{p}_{\|11}$

We optimize the likelihood using R function optim(), supplying a null vector (all 0) for the initial 7-parameter space, a maximum of 5000 iterations, gradient-based optimization method "L-BFGS-B", and we also include calculation of the Hessian matrix in order to compute standard error estimates for each parameter.

Table A1: Simulated bias of γ_2 ($\theta_y = e^{\gamma_2} = 1$, no treatment effect)

$\theta_x = 1$

			$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
		$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	$\phi_y = 2$	0.6	0.001	0.001	0.000	0.002	0.001	-0.002	-0.002	0.001	-0.002
		0.8	-0.002	-0.003	0.004	-0.003	0.001	0.002	-0.010	-0.004	0.002
		0.9	-0.005	-0.001	0.002	-0.005	-0.001	-0.003	0.005	-0.002	-0.004
	$\phi_y = 6$	0.6	-0.004	0.000	0.001	0.000	0.004	0.001	0.000	0.004	0.000
		0.8	0.002	0.001	0.002	-0.002	-0.001	0.000	0.004	-0.003	-0.003
		0.9	0.003	-0.005	0.000	-0.005	-0.001	-0.004	-0.002	-0.006	0.002
$\phi_x = 6$	$\phi_y = 2$	0.6	-0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.001	0.002
		0.8	0.001	-0.002	0.002	0.000	-0.001	0.002	-0.004	0.001	-0.002
		0.9	-0.001	-0.005	-0.003	-0.002	-0.003	0.000	-0.005	-0.001	-0.003
	$\phi_y = 6$	0.6	0.000	0.003	0.000	0.000	0.000	0.000	-0.001	0.001	-0.001
		0.8	0.001	-0.001	0.002	-0.003	-0.001	-0.002	-0.006	-0.004	-0.001
		0.9	-0.003	-0.003	0.005	-0.002	-0.002	0.000	-0.006	-0.003	-0.002

$\theta_x = \frac{5}{6}$

			$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
		$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	$\phi_y = 2$	0.6	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	-0.002	0.001	-0.001	0.001	0.001	-0.003	-0.001	-0.002	-0.001
		0.9	0.002	-0.002	0.003	0.001	0.002	0.001	0.004	-0.001	-0.002
	$\phi_y = 6$	0.6	-0.001	0.000	0.001	-0.001	0.000	0.000	-0.003	-0.001	-0.002
		0.8	-0.003	0.001	0.003	0.003	0.000	0.000	-0.003	-0.002	-0.001
		0.9	-0.002	0.003	0.000	-0.008	-0.001	-0.004	0.000	0.002	-0.003
$\phi_x = 6$	$\phi_y = 2$	0.6	0.000	0.000	0.001	-0.002	-0.001	0.002	0.000	0.001	0.002
		0.8	0.001	-0.003	0.002	-0.001	0.003	0.000	-0.002	-0.001	0.001
		0.9	0.005	-0.002	0.002	-0.001	0.000	0.004	-0.006	-0.002	-0.002
	$\phi_y = 6$	0.6	0.000	-0.001	-0.001	0.001	0.000	0.000	0.000	0.001	-0.003
		0.8	-0.001	-0.002	0.001	0.001	0.002	0.002	-0.004	0.001	0.000
		0.9	-0.002	0.001	-0.002	0.001	-0.001	0.000	-0.007	-0.006	0.004

$\theta_x = \frac{2}{3}$

			$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
		$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	$\phi_y = 2$	0.6	0.000	0.002	0.001	-0.002	0.000	-0.001	0.000	0.001	0.000
		0.8	0.003	0.002	0.001	-0.002	-0.002	0.001	-0.001	-0.003	-0.002
		0.9	0.000	0.000	0.003	-0.003	0.003	0.001	-0.006	-0.006	-0.003
	$\phi_y = 6$	0.6	-0.002	0.001	0.001	-0.003	0.000	-0.002	-0.001	0.000	0.000
		0.8	0.000	0.000	0.000	-0.001	-0.002	-0.002	-0.001	0.000	0.001
		0.9	-0.003	0.002	-0.002	0.003	-0.003	-0.002	-0.005	-0.004	-0.002
$\phi_x = 6$	$\phi_y = 2$	0.6	0.000	-0.001	-0.002	0.004	0.001	-0.001	-0.001	0.000	-0.001
		0.8	0.002	0.000	0.001	-0.001	0.004	-0.001	-0.008	0.001	0.001
		0.9	-0.002	-0.002	-0.001	0.001	0.002	0.002	-0.007	-0.008	-0.003
	$\phi_y = 6$	0.6	0.004	-0.003	-0.001	-0.002	0.000	-0.001	-0.001	0.001	-0.001
		0.8	-0.003	0.001	0.002	-0.003	-0.002	-0.002	-0.003	0.001	-0.001
		0.9	0.001	-0.006	-0.001	-0.001	-0.001	0.000	-0.007	-0.003	-0.002

Table A2: Simulated bias of γ_2 ($\theta_y = e^{\gamma_2} = \frac{5}{6}$, lower treatment effect)

$\theta_x = 1$

		$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
	$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	-0.001	-0.003	0.001	0.002	0.001	0.001	-0.001	0.002	-0.001
$\phi_y = 2$	0.8	-0.001	0.003	0.000	0.004	0.002	0.000	-0.004	-0.001	-0.001
	0.9	0.003	-0.001	0.001	-0.005	0.001	-0.002	0.002	-0.001	-0.007
$\phi_x = 2$	0.6	-0.001	0.000	-0.001	-0.001	0.002	0.000	0.001	-0.001	0.000
$\phi_y = 6$	0.8	-0.003	0.002	0.000	-0.001	-0.001	-0.002	-0.003	-0.001	0.000
	0.9	-0.001	0.000	-0.001	0.006	-0.001	-0.001	0.000	0.000	-0.002
$\phi_x = 6$	0.6	-0.003	0.000	-0.001	-0.004	-0.001	-0.002	-0.003	0.000	-0.001
$\phi_y = 2$	0.8	0.001	-0.003	0.002	-0.001	-0.003	0.000	-0.001	-0.002	-0.002
	0.9	-0.004	0.001	-0.003	-0.004	-0.001	0.003	-0.005	0.004	0.000
$\phi_x = 6$	0.6	-0.002	0.002	0.000	-0.001	0.000	0.000	0.000	0.001	0.000
$\phi_y = 6$	0.8	0.002	-0.001	0.000	0.001	-0.002	-0.001	0.002	-0.004	-0.002
	0.9	-0.002	0.001	-0.001	0.002	-0.004	-0.001	-0.005	-0.001	-0.004

$\theta_x = \frac{5}{6}$

		$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
	$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	0.000	-0.002	0.000	0.000	0.000	0.000	0.001	-0.002	-0.001
$\phi_y = 2$	0.8	-0.003	0.000	-0.001	-0.006	-0.002	-0.002	-0.006	0.000	-0.002
	0.9	-0.001	-0.004	0.001	0.000	-0.003	0.000	-0.005	-0.003	-0.001
$\phi_x = 2$	0.6	0.000	0.001	0.001	0.001	0.001	0.000	-0.002	0.001	-0.001
$\phi_y = 6$	0.8	-0.001	-0.002	-0.001	-0.001	-0.003	0.001	0.005	-0.002	-0.002
	0.9	0.001	-0.001	0.002	0.001	0.000	0.000	-0.009	-0.001	-0.001
$\phi_x = 6$	0.6	0.000	-0.001	0.000	-0.003	0.001	0.000	-0.004	0.002	0.001
$\phi_y = 2$	0.8	0.002	0.001	0.000	0.000	0.001	0.002	0.002	0.001	0.001
	0.9	0.002	-0.001	-0.002	-0.004	-0.002	0.001	-0.004	-0.002	-0.003
$\phi_x = 6$	0.6	0.002	0.000	0.000	0.000	0.001	0.001	-0.004	0.000	-0.002
$\phi_y = 6$	0.8	0.001	0.000	0.002	-0.003	-0.001	-0.001	0.000	-0.002	0.003
	0.9	-0.003	-0.001	0.001	0.002	-0.005	-0.001	-0.005	0.000	-0.001

$\theta_x = \frac{2}{3}$

		$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
	$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	-0.001	0.001	-0.001	0.000	-0.001	0.001	-0.002	0.002	0.001
$\phi_y = 2$	0.8	0.005	0.000	0.000	0.002	-0.002	0.001	-0.007	0.001	0.001
	0.9	-0.003	-0.003	-0.001	0.002	-0.006	-0.001	-0.008	0.000	0.000
$\phi_x = 2$	0.6	0.001	-0.002	0.001	-0.001	0.002	-0.001	0.000	0.000	-0.001
$\phi_y = 6$	0.8	0.000	-0.002	0.001	0.002	-0.002	-0.001	0.000	0.000	0.002
	0.9	-0.001	0.001	-0.005	-0.009	-0.003	-0.003	-0.001	0.001	-0.002
$\phi_x = 6$	0.6	-0.003	0.000	0.000	0.000	-0.001	0.001	0.000	-0.001	0.001
$\phi_y = 2$	0.8	-0.002	0.000	0.001	-0.004	0.001	-0.002	-0.003	-0.001	0.001
	0.9	-0.001	-0.001	-0.002	-0.003	-0.003	-0.002	-0.006	-0.002	-0.001
$\phi_x = 6$	0.6	0.000	0.003	0.001	0.000	-0.001	0.000	-0.001	0.001	0.000
$\phi_y = 6$	0.8	0.000	0.000	-0.002	0.001	-0.001	-0.002	-0.001	0.000	0.002
	0.9	-0.002	-0.003	-0.001	0.001	-0.002	-0.003	-0.002	-0.007	0.000

Table A3: Simulated bias of γ_2 ($\theta_y = e^{\gamma_2} = \frac{2}{3}$, higher treatment effect)

$\theta_x = 1$

		$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
	$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	-0.001	0.001	-0.001	0.001	0.001	0.002	-0.003	0.000	-0.001
$\phi_y = 2$	0.8	-0.001	0.000	0.000	-0.005	0.002	-0.003	-0.011	-0.004	-0.003
	0.9	0.001	-0.003	0.003	-0.003	-0.001	0.000	-0.003	-0.006	-0.004
$\phi_x = 2$	0.6	-0.001	-0.002	-0.002	-0.004	-0.001	0.000	0.000	0.000	-0.003
$\phi_y = 6$	0.8	0.000	-0.001	0.001	-0.003	0.000	0.000	-0.007	-0.001	0.002
	0.9	0.002	-0.002	0.001	0.001	-0.006	0.003	-0.003	-0.001	0.001

		0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	0.001	-0.001	-0.002	0.000	0.001	-0.001	0.000	-0.002	-0.001
$\phi_y = 2$	0.8	-0.001	-0.005	0.000	0.000	0.002	-0.005	-0.001	0.003	-0.002
	0.9	0.001	-0.002	0.000	-0.002	0.001	0.000	-0.008	-0.003	0.000
$\phi_x = 6$	0.6	0.001	0.000	-0.001	-0.002	0.000	0.002	0.005	0.003	-0.001
$\phi_y = 6$	0.8	0.001	0.000	0.001	-0.002	-0.001	0.002	-0.005	0.000	0.000
	0.9	-0.003	-0.001	0.001	-0.006	-0.002	-0.002	-0.006	-0.004	-0.002

$\theta_x = \frac{5}{6}$

		$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
	$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	-0.001	0.000	-0.003	-0.001	0.000	0.001	-0.001	0.002	-0.004
$\phi_y = 2$	0.8	0.001	0.001	-0.001	-0.001	0.002	0.000	-0.008	-0.002	-0.002
	0.9	-0.004	-0.002	0.000	0.003	-0.003	-0.001	-0.006	-0.004	-0.005
$\phi_x = 2$	0.6	0.001	-0.001	0.001	0.000	-0.001	-0.001	0.002	0.001	-0.002
$\phi_y = 6$	0.8	0.001	-0.002	-0.001	-0.003	0.002	-0.001	-0.004	0.002	0.001
	0.9	0.000	0.000	-0.001	-0.003	-0.002	0.000	0.001	-0.005	0.000

		0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	-0.001	0.001	0.002	-0.001	-0.001	-0.003	-0.003	0.001	-0.003
$\phi_y = 2$	0.8	0.001	0.001	0.003	-0.004	0.000	-0.001	-0.006	-0.002	0.001
	0.9	0.001	0.002	-0.004	-0.001	-0.001	-0.002	-0.012	0.001	0.002
$\phi_x = 6$	0.6	0.001	-0.001	0.001	-0.002	-0.002	0.000	0.002	-0.001	0.001
$\phi_y = 6$	0.8	-0.001	-0.001	0.001	0.001	-0.001	0.000	-0.004	-0.003	0.002
	0.9	-0.002	-0.001	-0.002	-0.003	-0.001	0.004	-0.001	-0.002	0.003

$\theta_x = \frac{2}{3}$

		$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
	$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	0.001	-0.002	-0.003	-0.002	0.000	0.001	0.000	0.002	-0.001
$\phi_y = 2$	0.8	-0.001	-0.001	-0.001	-0.002	-0.001	0.000	0.001	-0.001	0.002
	0.9	-0.004	-0.002	0.000	-0.006	0.000	0.001	0.002	-0.001	-0.003
$\phi_x = 2$	0.6	0.000	-0.002	0.001	-0.001	0.000	0.000	0.002	-0.001	-0.001
$\phi_y = 6$	0.8	-0.003	-0.001	-0.002	0.001	0.001	-0.001	-0.008	-0.004	0.000
	0.9	-0.002	-0.002	0.002	0.005	-0.004	0.003	-0.003	0.001	-0.005

		0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\phi_x = 2$	0.6	0.000	0.000	-0.001	0.000	0.001	0.000	-0.001	0.000	-0.002
$\phi_y = 2$	0.8	-0.001	-0.003	-0.003	0.000	-0.002	0.001	-0.005	-0.002	0.001
	0.9	0.000	0.000	0.001	-0.002	0.000	0.004	-0.003	-0.006	-0.006
$\phi_x = 6$	0.6	-0.002	0.000	0.000	-0.002	-0.002	0.000	0.000	0.001	0.001
$\phi_y = 6$	0.8	-0.003	-0.002	0.000	0.001	-0.002	-0.002	-0.004	0.001	0.001
	0.9	0.002	-0.001	-0.001	-0.001	-0.004	0.002	-0.006	-0.006	0.001

References

[1] S. J. Senn and R. A. Brown. Estimating treatment effects in clinical trials subject to regression to the mean. *Biometrics*, 41(2):555–560, 1985.

[2] S. Chinn and R. F. Heller. Some further results concerning regression to the mean. *American Journal of Epidemiology*, 114(6):902–905, 1981.

[3] K. J. Beath and A. J. Dobson. Regression to the mean for nonnormal populations. *Biometrika*, 78(2):431–435, 1991.

[4] M. John and A. F. Jawad. Assessing the regression to the mean for non-normal populations via kernel estimators. *N Am J Med Sci*, 2(7):288–92, 2010.

[5] H. G. Muller, I. Abramson, and R. Azari. Nonparametric regression to the mean. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17):9715–9720, 2003.

[6] Yuda Zhu and Robert E. Weiss. Modeling seroadaptation and sexual behavior among hiv+ study participants with a simultaneously multilevel and multivariate longitudinal count model. *Biometrics*, 69(1):214–224, 2013.

[7] J. P. Hughes, D. F. Haley, P. M. Frew, C. E. Golin, A. A. Adimora, I. Kuo, J. Justman, L. Soto-Torres, J. Wang, and S. Hodder. Regression to the mean and changes in risk behavior following study enrollment in a cohort of us women at risk for hiv. *Annals of Epidemiology*, 25(6):439–444, 2015.

[8] S. L. Hodder, J. Justman, J. P. Hughes, J. Wang, D. F. Haley, A. A. Adimora, C. Del Rio, C. E. Golin, I. Kuo, A. Rompalo, L. Soto-Torres, S. B. Mannheimer, L. Johnson-Lewis, S. H. Eshleman, W. M.

El-Sadr, and H. I. V. SeroIncidence Study Womens. Hiv acquisition among women from selected areas of the united states a cohort study. *Annals of Internal Medicine*, 158(1):10–U53, 2013.

6 Tables and Figures

Table 1: Power testing H_0 ($\gamma_2 = 0$)

			$\psi_c = 1.0$			$\psi_c = 1.5$			$\psi_c = 10$		
		$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\theta_y = \frac{5}{6}$	$\phi_y = 2$	0.6	0.860	0.964	0.982	0.849	0.958	0.984	0.858	0.968	0.987
		0.8	0.675	0.849	0.905	0.642	0.843	0.901	0.458	0.790	0.898
		0.9	0.404	0.578	0.659	0.364	0.562	0.636	0.157	0.380	0.565
	$\phi_y = 6$	0.6	0.906	0.976	0.991	0.904	0.976	0.991	0.868	0.979	0.991
		0.8	0.754	0.907	0.941	0.725	0.892	0.946	0.557	0.864	0.930
		0.9	0.487	0.655	0.722	0.427	0.635	0.714	0.220	0.483	0.659
		$p_y \backslash p_x$	0.6	0.8	0.9	0.6	0.8	0.9	0.6	0.8	0.9
$\theta_y = \frac{2}{3}$	$\phi_y = 2$	0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.8	1.000	1.000	1.000	1.000	1.000	1.000	0.996	1.000	1.000
		0.9	0.991	1.000	1.000	0.983	0.999	1.000	0.760	0.990	1.000
	$\phi_y = 6$	0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.9	0.997	1.000	1.000	0.995	1.000	1.000	0.882	0.999	1.000

Table 2: Bias of γ_2 when conditional odds ratio is not independent of study participation

($\psi_{c,off} \neq \psi_{c,on}$, $p_{x,off} = 0.75$, $\phi_x = 6.63$, $\theta_x = 0.96$ $p_{y,off} = 0.72$, $\phi_y = 2.83$)

		$\psi_{c,on}$									
		1.05	1.10	1.20	1.30	1.50	1.75	2.00	2.50	5.00	10.0
$\psi_{c,off}$	1.05	-	0.009	0.025	0.041	0.071	0.101	0.129	0.174	0.309	0.420
	1.10	-0.010	-	0.017	0.032	0.063	0.091	0.118	0.166	0.302	0.414
	1.20	-0.027	-0.016	-	0.016	0.045	0.076	0.103	0.149	0.286	0.400
	1.30	-0.043	-0.034	-0.017	-	0.029	0.059	0.088	0.135	0.273	0.387
	1.50	-0.07	-0.062	-0.045	-0.028	-	0.033	0.058	0.107	0.246	0.364
	1.75	-0.102	-0.095	-0.078	-0.061	-0.034	-	0.028	0.075	0.218	0.338
	2.00	-0.131	-0.123	-0.105	-0.087	-0.061	-0.028	-	0.048	0.193	0.315
	2.50	-0.182	-0.173	-0.156	-0.140	-0.109	-0.076	-0.049	-	0.149	0.273
	5.00	-0.349	-0.339	-0.321	-0.304	-0.275	-0.241	-0.211	-0.160	-	0.140
	10.0	-0.528	-0.517	-0.501	-0.483	-0.452	-0.415	-0.384	-0.33	-0.159	-

Table 3: Empirical and model-based prevalence estimates (%) among study population (n=2099)

	BL[a](missing)	final[b](missing)	on-study[c]	$\Delta p_{(BL-final)}$	$\Delta p_{(final-on)}$
any unprotected sex	95.7 (0.00)	77.0 (27.5[d]; 99.5[e])	73.9	18.7	3.1
last sex unprotected	81.9 (1.19)	63.0 (37.2; 73.6)	55.4	18.9	7.6
food insecurity	46.9 (1.29)	37.1 (28.0; 97.6)	35.4	9.8	1.7
foregone medical care	19.9 (0.00)	14.4 (27.4; 99.7)	13.6	5.5	0.8

[a] Baseline. [b] 12 months after study. [c] Estimated using eq. 5 [d] %-missing among full study population (n=2099) [e] Among missing data, %-missing due to loss of followup

Table 4: Primary model results for HPTN 064 study

Natural		Last Sex Unprotected			Food Insecurity			Foregone Medical Care		
Parameters	Univ.[a]	Est.	St. Err	95% C.I.	Est.	St. Err	95% C.I.	Est.	St. Err	95% C.I.
α_0	-0.17	-0.039	0.07	(-0.18, 0.11)	-0.17	0.08	(-0.32, -0.02)	-0.17	0.08	(-0.33, -0.02)
α_1	1.89	1.67	0.11	(1.45, 1.89)	1.89	0.12	(1.65, 2.12)	1.89	0.12	(1.66, 2.13)
α_2	-0.041	0.026	0.06	(-0.08, 0.14)	-0.039	0.06	(-0.15, 0.07)	-0.041	0.06	(-0.15, 0.07)
γ_0	-	0.31	0.07	(0.17, 0.44)	-0.88	0.05	(-0.97, -0.78)	-1.91	0.06	(-2.02, -1.79)
γ_1	-	1.04	0.08	(0.88, 1.20)	1.54	0.08	(1.39, 1.69)	1.81	0.10	(1.60, 2.01)
γ_2	-	-0.65	0.06	(-0.77, -0.54)	-0.33	0.05	(-0.43,-0.23)	-0.34	0.07	(-0.47, -0.21)
$\log(\psi_c)$	-	2.08	0.13	(1.83, 2.33)	0.14	0.09	(-0.05, 0.32)	0.042	0.13	(-0.21, 0.29)
Derived		Last Sex Unprotected			Food Insecurity			Foregone Medical Care		
parameters	Univ.	Est.	St. Err	95% C.I.	Est.	St. Err	95% C.I.	Est.	St. Err	95% C.I.
$p_{x,off}$	0.750	0.750	0.002	(0.745, 0.754)	0.750	0.002	(0.746, 0.754)	0.750	0.002	(0.746, 0.754)
$p_{x,on}$	0.739	0.756	0.012	(0.732, 0.781)	0.739	0.014	(0.712, 0.767)	0.739	0.015	(0.710, 0.768)
δ_x	0.050	-0.032	0.066	(-0.162, 0.098)	0.049	0.061	(-0.070, 0.168)	0.050	0.063	(-0.074, 0.174)
$p_{y,off}$	-	0.736	0.011	(0.714, 0.758)	0.463	0.012	(0.440, 0.486)	0.198	0.010	(0.179, 0.217)
$p_{y,on}$	-	0.554	0.012	(0.530, 0.579)	0.354	0.012	(0.330, 0.378)	0.136	0.008	(0.121, 0.151)
δ_y	-	0.687	0.028	(0.632, 0.741)	0.951	0.042	(0.869, 1.033)	0.983	0.056	(0.874, 1.091)

[a] Results from univariate model described in [7]