# Web Appendix for "Estimating the efficacy of pre-exposure prophylaxis for HIV prevention among participants with a threshold level of drug concentration"

We consider data from a two-phase sampling scheme for drug assays in PrEP trials. In the first phase, randomized assignment (Z) and HIV infection (Y) are collected for every participant. In the second phase, a case-control sample was drawn in the drug arm for measuring drug concentration (X). All three variables are binary indicators. Let $N_{ij}$ denote the number of participants with Z=$i$, and Y=$j$, $i, j$=0,1. Denote by $n_{ijk}$ the number of participants in the drug arm who are sampled into the second phase and have HIV infection status Y=$j$ and drug-level status X=$k$. Let $p_{ijk} = \Pr(\mathrm{Y} = j|\mathrm{X} = i, \mathrm{Z} = k)$, which can be expressed as a function of the parameters $(\beta_0, \beta_1, \beta_2)$ by logistic regression. Let $\alpha_k = \Pr(\mathrm{X} = k)$, $k = 0, 1$. The likelihood can be expressed as $\prod_j p_{0j}^{N_{0j}} p_{1j}^{N_{1j}-n_{1jk}} \prod_{j,k} (\alpha_k p_{1jk})^{n_{1jk}}$, where $p_{ij} = \Pr(\mathrm{Y} = j|\mathrm{Z} = i) = \sum_k p_{ijk}\alpha_k$. If X is observed for everyone, all parameters can be directly estimated. This leads to an EM algorithm that maximizes the likelihood. The fixed sensitivity parameter is included in the model by using an offset in the logistic regression inside the EM algorithm.

We simulated data with HIV infection (Y), drug assignment (Z), and drug presence (X) to examine the validity of our methods when we fix the sensitivity parameter $\delta$ in the estimation. We set $\beta_2 = \log(0.5)$ and $\beta_0 = -3.0$ so that HIV incidence is low and there is roughly 50% efficacy among participants who have a detectable level of drug. We vary $\beta_1$ from 0, where there is no difference between baseline HIV risk of high adherer and low adherer, to $\log 1.5$, the high adherer has a lower HIV risk compared to the low adherer. Case-control sampling is employed in the active treatment arm with a 1:1 sampling of infected:uninfected subjects for applying the drug assay. The frequency of drug compliers is 60%. Web Table 1 shows the parameter estimates of $\beta_1$ and $\beta_2$ for $\delta = (0, \log 0.9, \log 0.8)$ with a sample size of 2500 or 5000. The results suggest that the estimation procedure works fine. Bias is generally small and the estimated variance is mostly close to, if not greater than, the empirical variability of the estimator, leading to slightly conservative confidence intervals.

Web Table 1: Simulations to Assess the Performance of the Estimation Method [a].

| | $\hat{\beta}_1$ | | | | $\hat{\beta}_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | Var | $\widehat{\text{Var}}$ | 95% CP | Bias | Var | $\widehat{\text{Var}}$ | 95% CP |
| $\beta_1 = 0, \quad N = 2500$ | | | | | | | | |
| $\delta = 0$ | −0.018 | 0.289 | 0.301 | 0.970 | 0.005 | 0.145 | 0.146 | 0.967 |
| $\delta = \log 0.9$ | −0.026 | 0.334 | 0.328 | 0.963 | 0.001 | 0.160 | 0.157 | 0.968 |
| $\delta = \log 0.8$ | −0.031 | 0.368 | 0.362 | 0.968 | −0.002 | 0.168 | 0.169 | 0.969 |
| $\beta_1 = -\log 1.5, \quad N = 2500$ | | | | | | | | |
| $\delta = 0$ | 0.009 | 0.352 | 0.429 | 0.968 | −0.028 | 0.209 | 0.256 | 0.966 |
| $\delta = \log 0.9$ | 0.004 | 0.380 | 0.461 | 0.966 | −0.040 | 0.226 | 0.272 | 0.966 |
| $\delta = \log 0.8$ | 0.010 | 0.423 | 0.502 | 0.964 | −0.055 | 0.240 | 0.291 | 0.969 |
| $\beta_1 = 0, \quad N = 5000$ | | | | | | | | |
| $\delta = 0$ | −0.031 | 0.148 | 0.141 | 0.943 | 0.013 | 0.062 | 0.067 | 0.971 |
| $\delta = \log 0.9$ | −0.007 | 0.158 | 0.154 | 0.951 | 0.010 | 0.068 | 0.071 | 0.968 |
| $\delta = \log 0.8$ | −0.008 | 0.164 | 0.169 | 0.968 | 0.007 | 0.073 | 0.077 | 0.964 |
| $\beta_1 = -\log 1.5, \quad N = 5000$ | | | | | | | | |
| $\delta = 0$ | −0.028 | 0.204 | 0.205 | 0.959 | 0.016 | 0.109 | 0.121 | 0.971 |
| $\delta = \log 0.9$ | −0.017 | 0.207 | 0.221 | 0.963 | 0.009 | 0.112 | 0.128 | 0.972 |
| $\delta = \log 0.8$ | −0.031 | 0.223 | 0.244 | 0.969 | 0.006 | 0.127 | 0.140 | 0.968 |

Abbreviations: Var, the empirical variance; $\widehat{\text{Var}}$, the average of the estimated variances; 95% CP - coverage probability of 95% confidence interval.

[a] HIV infection data are generated based on $\text{logit}\Pr(Y = 1) = -3 + \delta Z + \beta_1 X + \log 0.5 XZ$ for sample size 2500 or 5000.