Supplemental material

Technical details for bivariate linkage model

Maximum likelihood estimation

Combining (3) and (4), the distribution for $(\widehat{U}_m, \widehat{V}_m)$ is given by

$$\begin{pmatrix}\widehat{U}_m\\ \widehat{V}_m\end{pmatrix} \sim MVN_2\left(\begin{pmatrix}\mu_U\\ \mu_V\end{pmatrix}, \begin{pmatrix}\Sigma_{11}^{\widehat{U}_m,\widehat{V}_m} & \Sigma_{12}^{\widehat{U}_m,\widehat{V}_m}\\ \Sigma_{21}^{\widehat{U}_m,\widehat{V}_m} & \Sigma_{22}^{\widehat{U}_m,\widehat{V}_m}\end{pmatrix}\right), \tag{10}$$

where $\Sigma_{11}^{\widehat{U}_m,\widehat{V}_m} = \sigma_U^2 + s_{U,m}^2$, $\Sigma_{22}^{\widehat{U}_m,\widehat{V}_m} = \sigma_V^2 + s_{V,m}^2$, $\Sigma_{12}^{\widehat{U}_m,\widehat{V}_m} = \Sigma_{21}^{\widehat{U}_m,\widehat{V}_m} = \rho_m^*\sqrt{(\sigma_U^2 + s_{U,m}^2)(\sigma_V^2 + s_{V,m}^2)}$, and $\rho_m^* = \rho\sqrt{\sigma_U^2\sigma_V^2/[(\sigma_U^2 + s_{U,m}^2)(\sigma_V^2 + s_{V,m}^2)]} < \rho$. That is, the association between $\widehat{U}_m$ and $\widehat{V}_m$ is weaker than that between U_m and V_m; it is attenuated given the additional variability of $(\widehat{U}_m, \widehat{V}_m)$ given (U_m, V_m). The log-likelihood function for observed HIV and marker incidence $(\widehat{U}_m, \widehat{V}_m)$ for external cohort $m = 1, \ldots, M$ is then given by

$$\begin{aligned} l_M = &-M\log(2\pi) - \frac{1}{2}\sum_{m=1}^{M}\log\left\{(\sigma_U^2 + s_{U,m}^2)(\sigma_V^2 + s_{V,m}^2) - \rho^2\sigma_U^2\sigma_V^2\right\}\\ &-\sum_{m=1}^{M}\frac{(\sigma_V^2 + s_{V,m}^2)\left(\widehat{U}_m - \mu_U\right)^2 - 2\rho\sqrt{\sigma_U^2\sigma_V^2}\left(\widehat{U}_m - \mu_U\right)\left(\widehat{V}_m - \mu_V\right) + (\sigma_U^2 + s_{U,m}^2)\left(\widehat{V}_m - \mu_V\right)^2}{2\left\{(\sigma_U^2 + s_{U,m}^2)(\sigma_V^2 + s_{V,m}^2) - \rho^2\sigma_U^2\sigma_V^2\right\}}. \end{aligned}$$

We consider an EM algorithm treating (U_m, V_m) as missing data. Specifically, in the M-step, the log-likelihood for (U_m, V_m) ($m = 1, \ldots, M$) is given by

$$\begin{aligned} l_M = &-M\log(2\pi\sigma_U\sigma_V\sqrt{1-\rho^2})\\ &-\frac{1}{2(1-\rho^2)}\sum_{m=1}^{M}\left\{\left(\frac{U_m - \mu_U}{\sigma_U}\right)^2 - 2\rho\left(\frac{U_m - \mu_U}{\sigma_U}\right)\left(\frac{V_m - \mu_V}{\sigma_V}\right) + \left(\frac{V_m - \mu_V}{\sigma_V}\right)^2\right\}. \end{aligned}$$

At the jth iteration, the maximizer of $(\mu_U, \mu_V, \sigma_U^2, \sigma_V^2, \rho)$ given (U_m, V_m) is given by

$$\begin{aligned} \widehat{\mu}_U^{(j)} =& M^{-1}\sum_{m=1}^{M}U_m, \quad \widehat{\mu}_V^{(j)} = M^{-1}\sum_{m=1}^{M}V_m,\\ \widehat{\sigma}_U^{2,(j)} =& M^{-1}\sum_{m=1}^{M}(U_m - \widehat{\mu}_U^{(j)})^2, \quad \widehat{\sigma}_V^{2,(j)} = M^{-1}\sum_{m=1}^{M}(V_m - \widehat{\mu}_V^{(j)})^2,\\ \widehat{\rho}^{(j)} =& \frac{M^{-1}\sum_{m=1}^{M}(U_m - \widehat{\mu}_U^{(j)})(V_m - \widehat{\mu}_V^{(j)})}{\widehat{\sigma}_U^{(j)}\widehat{\sigma}_V^{(j)}}. \end{aligned}$$

In the E-step, we evaluate the conditional expectation of the terms $(U_m, V_m, U_m^2, V_m^2, U_mV_m)$ given the observed data $(\widehat{U}_m, \widehat{V}_m)$ and current parameter values $(\widehat{\mu}_U^{(j)}, \mu_V^{(j)}, \sigma_U^{2,(j)}, \sigma_V^{2,(j)}, \rho^{(j)})$. Note that the joint distribution of $(\widehat{U}_m, \widehat{V}_m, U_m, V_m)^{\mathrm{T}}$ is given by a multivariate normal, with mean $(\mu_U, \mu_V, \mu_U, \mu_V)^{\mathrm{T}}$ and the covariance matrix is given by

$$\begin{pmatrix} \sigma_U^2 & \rho\sigma_U\sigma_V & \sigma_U^2 & \rho\sigma_U\sigma_V\\ \rho\sigma_U\sigma_V & \sigma_V^2 & \rho\sigma_U\sigma_V & \sigma_V^2\\ \sigma_U^2 & \rho\sigma_U\sigma_V & \sigma_U^2 + s_{U,m}^2 & \rho\sigma_U\sigma_V\\ \rho\sigma_U\sigma_V & \sigma_V^2 & \rho\sigma_U\sigma_V & \sigma_V^2 + s_{V,m}^2 \end{pmatrix}.$$

Then, the conditional distribution of $(U_m, V_m)^{\mathrm{T}}$ given $(\widehat{U}_m, \widehat{V}_m)^{\mathrm{T}}$ is bivariate normal, with mean

$$\begin{pmatrix}\widetilde{\mu}_{U,m}\\ \widetilde{\mu}_{V,m}\end{pmatrix} = \begin{pmatrix}\mu_U\\ \mu_V\end{pmatrix} + AB_m^{-1}\begin{pmatrix}\widehat{U}_m - \mu_U\\ \widehat{V}_m - \mu_V\end{pmatrix}$$

and variance matrix

$$\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{pmatrix} = A - AB_m^{-1}A,$$

where

$$A = \begin{pmatrix}\sigma_U^2 & \rho\sigma_U\sigma_V\\ \rho\sigma_U\sigma_V & \sigma_V^2\end{pmatrix}, \quad B_m = \begin{pmatrix}\sigma_U^2 + s_{U,m}^2 & \rho\sigma_U\sigma_V\\ \rho\sigma_U\sigma_V & \sigma_V^2 + s_{V,m}^2\end{pmatrix}.$$

Based on the proceeding derivations, for each step of the EM algorithm, we updated $(\mu_U, \mu_V, \sigma_U^2, \sigma_V^2, \rho)$ by

$$\begin{aligned}
\widehat{\mu}_U^{(j)} =& M^{-1}\sum_{m=1}^{M}\widetilde{\mu}_{U,m}, \quad \widehat{\mu}_V^{(j)} = M^{-1}\sum_{m=1}^{M}\widetilde{\mu}_{V,m} \\
\widehat{\sigma}_U^{2,(j)} =& M^{-1}\sum_{m=1}^{M}(\widetilde{\mu}_{U,m}^2 + \Sigma_{11} - 2*\widehat{\mu}_U^{(j)}\widetilde{\mu}_{U,m} + (\widehat{\mu}_U^{(j)})^2), \\
\widehat{\sigma}_V^{2,(j)} =& M^{-1}\sum_{m=1}^{M}(\widetilde{\mu}_{V,m}^2 + \Sigma_{22} - 2*\widehat{\mu}_V^{(j)}\widetilde{\mu}_{V,m} + (\widehat{\mu}_V^{(j)})^2), \\
\widehat{\rho}^{(j)} =& \frac{M^{-1}\sum_{m=1}^{M}\widetilde{\mu}_{U,m}\widetilde{\mu}_{V,m} + \Sigma_{12} - \widehat{\mu}_U^{(j)}\widetilde{\mu}_{V,m} - \widehat{\mu}_V^{(j)}\widetilde{\mu}_{U,m} + \widehat{\mu}_U^{(j)}\widehat{\mu}_V^{(j)}}{\widehat{\sigma}_U^{(j)}\widehat{\sigma}_V^{(j)}},
\end{aligned}$$

where $\widetilde{\mu}_{U,m}, \widetilde{\mu}_{V,m}, \Sigma_{11}, \Sigma_{12}$, and Σ_{22} are evaluated at the parameter value at the last step $(\widehat{\mu}_U^{(j-1)}, \widehat{\mu}_V^{(j-1)}, \widehat{\sigma}_U^{2,(j-1)}, \widehat{\sigma}_V^{2,(j-1)}, \widehat{\rho}^{(j-1)})$. We iterative between the E-step and M-step till the algorithm converges.

Let $\widehat{\Sigma}$ be the estimated covariance matrix of $(\widehat{\mu}_U, \widehat{\mu}_V, \widehat{\sigma}_U^2, \widehat{\sigma}_V^2, \widehat{\rho})$ based on the maximum likelihood approach. By the delta method, the variance of $\widehat{U}_{0,k}$ can be estimated by

$$\widehat{var}(\widehat{U}_{0,k}) = c_1 \begin{pmatrix} \widehat{\Sigma} & 0 \\ 0 & \widehat{s}_{V,k}^2 \end{pmatrix} c_1^{\mathrm{T}},$$

where

$$c_1 = \left(1, -\frac{\widehat{\rho}\widehat{\sigma}_U\widehat{\sigma}_V}{\widehat{\sigma}_V^2 + \widehat{s}_{V,k}^2}, \frac{\widehat{\rho}\widehat{\sigma}_V}{2\widehat{\sigma}_U(\widehat{\sigma}_V^2 + \widehat{s}_{V,k}^2)}(\widehat{V}_k - \widehat{\mu}_V), \frac{\widehat{\rho}\widehat{\sigma}_U(\widehat{s}_{V,k}^2 - \widehat{\sigma}_V^2)}{2\widehat{\sigma}_V(\widehat{\sigma}_V^2 + \widehat{s}_{V,k}^2)^2}(\widehat{V}_k - \widehat{\mu}_V), \frac{\widehat{\rho}\widehat{\sigma}_U\widehat{\sigma}_V}{\widehat{\sigma}_V^2 + \widehat{s}_{V,k}^2}\right),$$

such that the variance of $\widehat{\lambda}_{0,k}^Y$ can be estimated by

$$\widehat{var}(\widehat{\lambda}_{0,k}^Y) = \{\widehat{\lambda}_{0,k}^Y\}^2\widehat{var}(\widehat{U}_{0,k})$$

with log-link, or

$$\widehat{var}(\widehat{\lambda}_{0,k}^Y) = \{\widehat{\lambda}_{0,k}^Y(1-\widehat{\lambda}_{0,k}^Y)\}^2\widehat{var}(\widehat{U}_{0,k})$$

with logit-link.

By applying the delta method again. Then, the variance of $\log(\widehat{\lambda}_k^Y/\widehat{\lambda}_{0,k}^Y) = \widehat{U}_k - \widehat{U}_{0,k}$ can be estimated by

$$\frac{1-\widehat{\lambda}_k^Y}{n_{event,k}} + \widehat{var}(\widehat{U}_{0,k}),$$

with log-link, and the variance of $\widehat{PE}_k$ can be estimated by

$$\left(-\frac{1}{\widehat{\lambda}_{0,k}^Y}\right)^2\widehat{var}(\widehat{\lambda}_k^Y) + \left(\frac{\widehat{\lambda}_k^Y}{(\widehat{\lambda}_{0,k}^Y)^2}\right)^2\widehat{var}(\widehat{\lambda}_{0,k}^Y) = \frac{1}{(\widehat{\lambda}_{0,k}^Y)^2 n_{event,k}} + \frac{(1-\widehat{\lambda}_{0,k}^Y)^2\widehat{var}(\widehat{U}_{0,k})}{(\widehat{\lambda}_{0,k}^Y)^2},$$

with logit-link, where $n_{event,k}$ is the total number of events observed in the arm k in the trial.

Working regression model

The conditional distribution of $\widehat{U}_m$ given $\widehat{V}_m$ is given by

$$\widehat{U}_m = \alpha_m + \beta_m\widehat{V}_m + \widehat{\epsilon}_m, \tag{11}$$

where $\alpha_m = \mu_U - \rho\frac{\sigma_U\sigma_V}{\sigma_V^2+s_{V,m}^2}\mu_V$, $\beta_m = \rho\frac{\sigma_U\sigma_V}{\sigma_V^2+s_{V,m}^2}$, and $\widehat{\epsilon}_m \sim N(0, (\sigma_U^2 + s_{U,m}^2)(1-{\rho_m^*}^2))$. Note that the coefficients α_m and β_m are cohort-specific. This suggests that naively fitting (7) may not produce an unbiased estimate of the regression function, f. Indeed, (7) is correctly specified if and only if $s_{U,m}^2$ and $s_{V,m}^2$ are the same for all $m = 1, \ldots, M$.

The CI for $\widehat{\lambda}_{0,k}^{*,Y}$ is constructed based on a t-distribution approximation, given an asymptotic variance derived using the delta method. The analytical variance of $\widehat{PE}_k^*$ is non-trivial, as it involves the ratio between asymptotically normally-distributed $\widehat{\lambda}_k^Y$ and approximately t-distributed $\widehat{\lambda}_{0,k}^{*,Y}$, see e.g.[46,47] Therefore, to quantify uncertainty in $\widehat{PE}_k^*$ we apply the bootstrap whereby M external cohorts are sampled with replacement. Specifically, in each bootstrap iteration, we randomly sample M external cohorts with replacement and fit the working regression model based on the sampled M external cohorts. We also sample individuals with replacement from the trial to construct an estimate of the exposure marker $\widehat{\lambda}_Z$ and $\widehat{\lambda}_Y$. We then estimate the counterfactual placebo incidence and prevention efficacy by applying the bootstrapped incidence estimates to the bootstrapped working regression model.

Additional simulation studies

Impact of violation of the conditional independence assumption

We performed additional simulations in which the conditional independence assumption is violated. We assume the joint distribution of $(\widehat{U}_m, \widehat{V}_m)$ conditional on (U_m, V_m) has the following form:

$$\begin{pmatrix}\widehat{U}_m\\ \widehat{V}_m\end{pmatrix}\Bigg|\begin{pmatrix}U_m\\ V_m\end{pmatrix} \sim MVN_2\left(\begin{pmatrix}U_m\\ V_m\end{pmatrix}, \begin{bmatrix}s^2_{U,m} & \rho_m\\ \rho_m & s^2_{V,m}\end{bmatrix}\right), \tag{12}$$

where ρ_m measures the cohort-specific association between $\widehat{U}_m$ and $\widehat{V}_m$. The external cohort estimates were generated from (12), with ρ_m randomly generated from $U(0.4, 0.5)$. We implemented likelihood-based and working model estimation methods that incorrectly assume conditional independence. As shown in Supplementary Tables S5 and S6, the performance of the counterfactual placebo incidence and PE estimates is similar to that under a conditional independence model, suggesting that the estimation is reasonably robust to violation of the conditional independence assumption. Similar robustness has been reported in related bivariate meta-analysis work.[41] One of the conditions that minimizes the impact of conditional dependence is small within-study variation relative to between-study variation, e.g. due to large external cohort sizes. This condition is largely satisfied under our simulation scenarios.

External cohorts: more cohorts of smaller sizes or fewer cohorts of larger sizes?

In settings where external cohort data are available at a more granular level, e.g., when site-level data are available for multi-center studies, a question is whether performance of the counterfactual placebo estimate can be improved when the external cohort data are analyzed at the sub-cohort level. We conducted additional simulations for exploration. In these simulations, each of $M = 20$ cohorts has $L = 5$ sites and site-level HIV and HIV exposure marker incidences follow model (10). Person-years for each site is uniformly distributed as $U(200/L, 5000/L) = U(40, 1000)$. We compared results based on estimation of counterfactual placebo HIV incidence using cohort-level vs. site-level data, where the total sample size of the external cohort data is approximately constant. We found that analyzing data at the site level provides more precise inference, as reflected by narrower CIs with similar coverage rates (see Supplementary Table S7). Thus, given a fixed total sample size across external cohorts, more cohorts of smaller sizes is preferred to fewer cohorts of larger size.

Supplementary Tables and Figures

Table S1. Estimated HIV and rectal gonorrhea incidence rates for studies used as basis for simulations and analyzed in Mulick and Murray[24]

Referenced Study	HIV inc. cases/100 person-years	Num. person-years	RGC inc. cases/100 person-years	Num. person-years
Morris et al.[48]	2.5	943.2	3.5	943.2
Jin et al.[49]	0.9	5160*	2.3	5160*
Molina et al.[50]	6.6	212.1	15.5	212.1
Castillo et al.[51]	3.6	1000*	10.1	1000*
Kelley et al.[52]	3.8	843.1	6.2	726.6
McGowan et al.[53]	6.4	50*	16.1	50*
McCormack et al.[54]	9.0	245	33.1	596
Girometti et al.[55]	8.3	100*	33.0	100*

∗: person-years are approximate and not explicitly reported in the referenced studies.

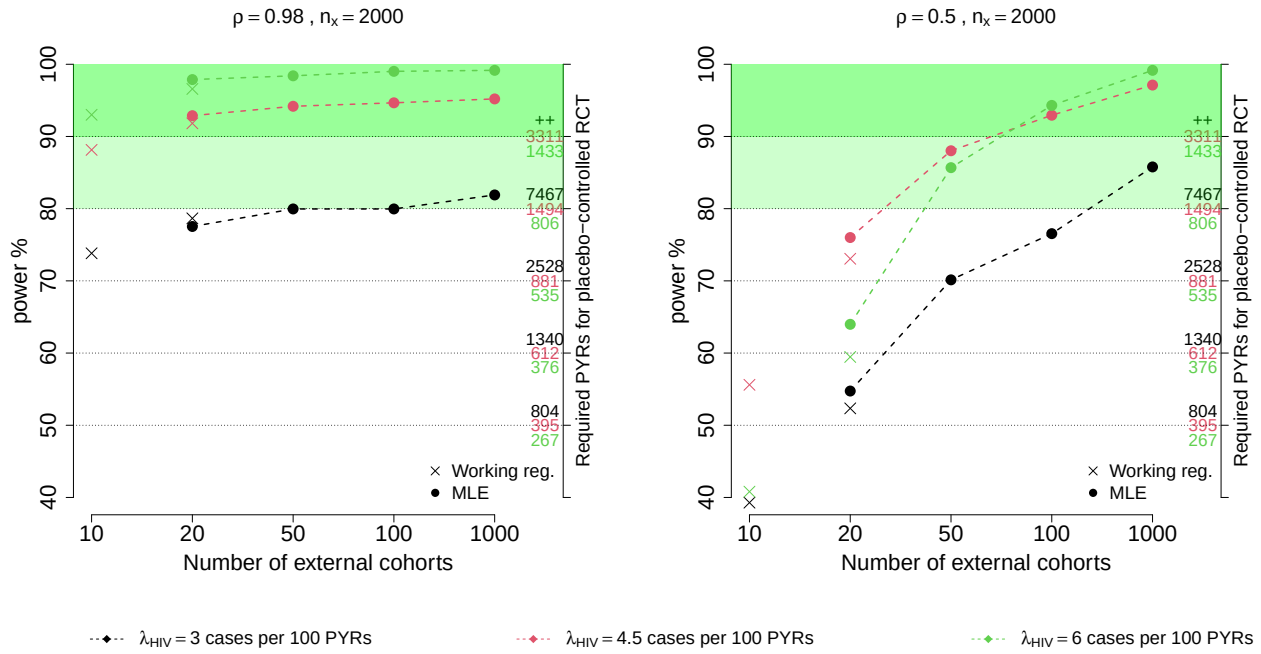

Figure S1. Power for testing $H_0: \ PE = 0.3$ vs. $H_a: \ PE = 0.6$ using the counterfactual approach as a function of M, the number of external cohorts. Given a fixed active arm size of 2000 person-years, the power based on estimating counterfactual placebo HIV incidence with a highly correlated marker ($\rho = 0.98$; left) or a moderately correlated exposure marker ($\rho = 0.5$; right) are shown. The size (person-years) of a placebo arm required to obtain the power shown on the left hand y-axis is shown on the right-hand y-axis. Reflecting performance of the two estimation approaches, power is shown for the working model approach for $M = 10, 20$ and for maximum likelihood (MLE) estimation with $M = 20, 50, 100, 1000$.

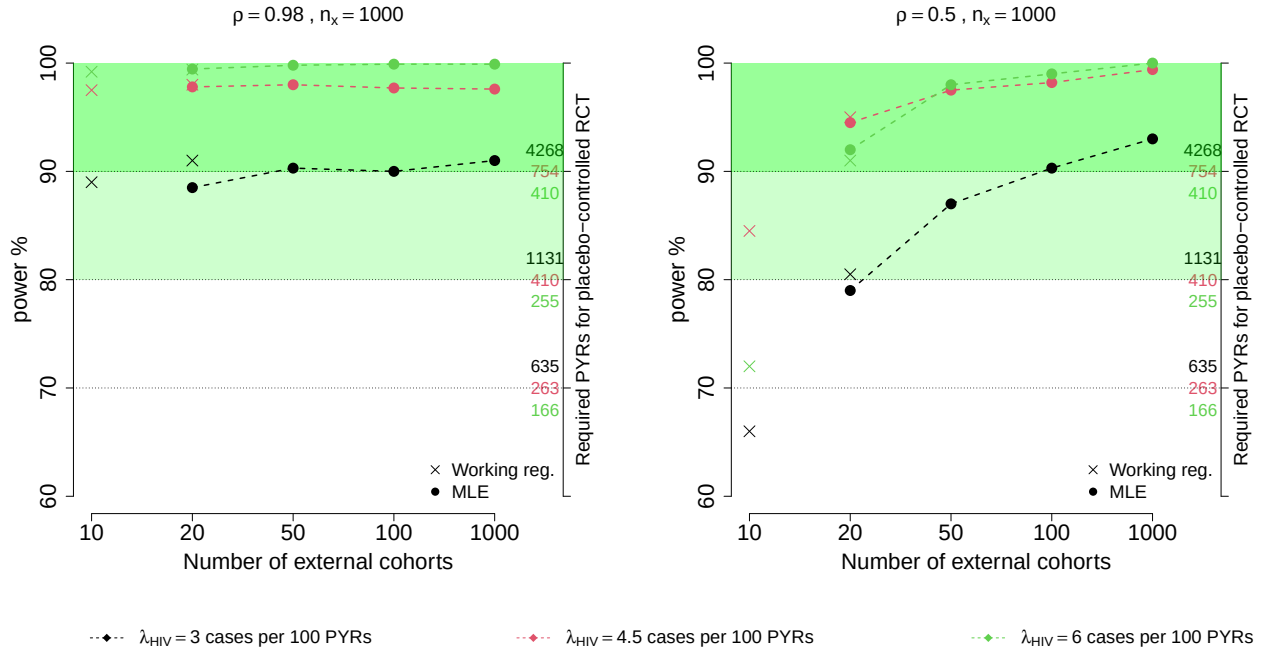

Figure S2. Power for testing $H_0: \ PE = 0.3$ vs. $H_a: \ PE = 0.75$ using the counterfactual approach as a function of M, the number of external cohorts. Given a fixed active arm size of 1000 person-years, the power based on estimating counterfactual placebo HIV incidence with a highly correlated marker ($\rho = 0.98$; left) or a moderately correlated exposure marker ($\rho = 0.5$; right) are shown. The sizes (person-years) of a placebo arm, required to obtain the power shown on the left hand y-axis is shown on the right-hand y-axis. Reflecting performance of the two estimation approaches, power is shown for the working model approach for $M = 10, 20$ and for maximum likelihood (MLE) estimation with $M = 20, 50, 100, 1000$.

Table S2. Bias, standard deviation, and empirical coverage for estimated counterfactual placebo HIV incidence, based on M external cohorts used to estimate the association between HIV and an exposure biomarker with correlation ρ. A total of n_x person-years follow-up accrue in the active arm of the trial. Counterfactual placebo HIV incidence varies. Performance is shown for working model and likelihood-based estimation approaches, assuming the logit link function for marginal incidences.

			$\rho = 0.971$			$\rho = 0.5$		
HIV incidence (cases per 100 person-years)			3	4.5	6	3	4.5	6
Exposure marker incidence (cases per 100 person-years)			7.2	12.3	17.6	4.8	13.5	26.2
			Working model approach					
$M = 10$	$n_x = 2000$	Bias × 100	-0.01	-0.02	-0.03	0.12	0.07	0.19
		Standard deviation × 100	0.35	0.41	0.56	1.04	1.05	2.09
		Coverage (%)	95.7	97.3	96.7	95.2	95.2	95.0
	$n_x = 4000$	Bias × 100	0.00	-0.02	-0.02	0.14	0.06	0.27
		Standard deviation × 100	0.33	0.38	0.53	1.05	1.02	2.14
		Coverage (%)	94.9	96.3	97.0	95.0	95.9	95.4
$M = 20$	$n_x = 2000$	Bias × 100	-0.02	-0.03	-0.03	0.05	0.01	0.06
		Standard deviation × 100	0.27	0.32	0.41	0.65	0.70	1.28
		Coverage (%)	94.8	96.2	97.0	95.6	95.4	95.4
	$n_x = 4000$	Bias × 100	-0.01	-0.03	-0.04	0.05	0.02	0.07
		Standard deviation × 100	0.24	0.28	0.37	0.65	0.71	1.34
		Coverage (%)	94.5	95.7	97.1	94.9	94.6	94.9
			Likelihood-based approach					
$M = 20$	$n_x = 2000$	Bias × 100	0.03	0.02	0.02	0.11	0.09	0.16
		Standard deviation × 100	0.26	0.31	0.41	0.66	0.71	1.33
		Coverage (%)	95.3	95.2	94.8	93.9	93.9	93.8
	$n_x = 4000$	Bias × 100	0.02	0.03	0.03	0.11	0.09	0.14
		Standard deviation × 100	0.22	0.28	0.36	0.65	0.69	1.31
		Coverage (%)	95.3	94.2	94.6	94.1	93.8	94.3

Table S3. Bias, standard deviation, and empirical coverage for estimates of prevention efficacy (PE) based on M external cohorts used to estimate the association between HIV and an exposure biomarker with correlation ρ. A total of $n_x = 2000$ person-years follow-up accrue in the active arm of the trial. Counterfactual placebo HIV incidence and true PE vary. Performance is shown for working model and likelihood-based estimation approaches, assuming logit link function for marginal incidences.

	M	PE estimate	$\rho = 0.971$			$\rho = 0.5$		
HIV incidence (cases per 100 person-years)			3	4.5	6	3	4.5	6
Exposure marker incidence (cases per 100 person-years)			7.2	12.3	17.6	4.8	13.5	26.2
			Working model approach					
$PE = 0.3$	10	Bias × 100	-1.08	-0.62	-0.85	-4.62	-2.56	-5.41
$n_x = 2000$		Standard deviation × 100	14.76	11.44	10.62	30.04	19.23	29.27
		Coverage (%)	92.8	93.3	93.2	91.7	92.2	90.2
	20	Bias × 100	-0.88	-0.75	-0.67	-1.84	-1.38	-2.10
		Standard deviation × 100	13.59	10.58	9.60	20.09	14.28	17.27
		Coverage (%)	92.8	93.3	93.0	93.0	93.8	93.1
$PE = 0.6$	10	Bias × 100	-0.51	-0.34	-0.51	-2.02	-1.12	-3.07
$n_x = 2000$		Standard deviation × 100	9.83	7.86	7.38	17.88	11.83	17.19
		Coverage (%)	93.9	94.1	93.6	92.4	93.4	90.9
	20	Bias × 100	-0.39	-0.55	-0.22	-0.59	-0.83	-1.22
		Standard deviation × 100	9.45	7.59	6.67	12.64	9.44	11.01
		Coverage (%)	93.5	93.5	93.9	93.7	94.1	92.6
$PE = 0.75$	10	Bias × 100	-0.21	-0.21	-0.22	-1.10	-0.53	-1.76
$n_x = 2000$		Standard deviation × 100	7.34	6.01	5.29	12.17	8.08	11.58
		Coverage (%)	93.9	93.8	94.5	92.9	93.2	92.2
	20	Bias × 100	-0.14	-0.13	-0.19	-0.16	-0.37	-0.73
		Standard deviation × 100	7.14	5.84	4.97	8.88	6.66	7.53
		Coverage (%)	93.3	94.0	94.2	93.7	94.2	93.3
			Likelihood-based approach					
$PE = 0.3$	20	Bias × 100	-0.19	-0.03	-0.19	-0.78	-0.21	-1.31
$n_x = 2000$		Standard deviation × 100	12.48	9.97	8.77	18.78	13.95	17.34
		Coverage (%)	93.9	94.0	94.7	91.6	93.1	92.0
$PE = 0.6$	20	Bias × 100	0.07	0.02	-0.01	-0.42	0.03	-0.54
$n_x = 2000$		Standard deviation × 100	8.81	7.07	6.42	11.73	9.25	10.46
		Coverage (%)	94.0	94.4	93.5	92.1	92.5	92.3
$PE = 0.75$	20	Bias × 100	0.04	0.00	0.12	-0.26	-0.05	-0.56
$n_x = 2000$		Standard deviation × 100	6.79	5.50	4.83	8.33	6.52	7.20
		Coverage (%)	93.3	93.8	94.2	91.8	92.6	93.2

Table S4. Bias, standard deviation, and empirical coverage for estimates of prevention efficacy (PE) based on M external cohorts used to estimate the association between HIV and an exposure biomarker with correlation ρ. A total of $n_x = 1000$ person-years follow-up accrue in the active arm of the trial. Counterfactual placebo HIV incidence and true PE vary. Performance is shown for working model and likelihood-based estimation approaches, assuming log link function for marginal incidences.

	M	PE estimate	$\rho = 0.98$			$\rho = 0.5$		
HIV incidence (cases per 100 person-years)			3	4.5	6	3	4.5	6
Exposure marker incidence (cases per 100 person-years)			7.1	11.8	17.0	4.8	13.2	26.7
			Working model approach					
$PE = 0.6$	10	Bias × 100	-0.66	-0.72	-0.54	-2.28	-0.58	-2.18
$n_x = 1000$		Standard deviation × 100	13.59	10.79	9.49	20.57	14.17	17.66
		Coverage (%)	93.1	93.8	93.7	93.6	93.1	92.4
	20	Bias × 100	-0.68	-0.61	-0.36	-1.13	-0.36	-0.86
		Standard deviation × 100	13.49	10.6	9.18	15.83	11.77	12.72
		Coverage (%)	93.0	93.4	93.5	93.9	93.9	93.3
$PE = 0.75$	10	Bias × 100	-0.18	-0.30	-0.23	-1.17	-0.14	-1.13
$n_x = 1000$		Standard deviation × 100	10.24	8.18	7.18	14.31	10.07	11.84
		Coverage (%)	92.8	93.8	93.7	93.3	93.0	92.9
	20	Bias × 100	-0.22	-0.28	-0.12	-0.47	-0.07	-0.36
		Standard deviation × 100	10.25	8.12	7.02	11.52	8.75	8.95
		Coverage (%)	92.8	93.4	93.5	93.5	93.3	93.6
			Likelihood-based approach					
$PE = 0.6$	20	Bias × 100	-0.10	0.01	-0.05	-0.11	-0.45	-0.63
$n_x = 1000$		Standard deviation × 100	12.34	9.96	8.72	14.42	11.66	12.34
		Coverage (%)	95.7	95.4	94.9	95.2	94.4	93.8
$PE = 0.75$	20	Bias × 100	-0.07	0.00	-0.04	-0.06	-0.30	-0.41
$n_x = 1000$		Standard deviation × 100	9.56	7.75	6.77	10.61	8.69	8.76
		Coverage (%)	96.5	95.6	94.9	96.2	94.9	94.1

Table S5. Bias, standard deviation, and empirical coverage for estimated counterfactual placebo HIV incidence, based on M external cohorts used to estimate the association between HIV and an exposure biomarker with correlation ρ. A total of n_x person-years follow-up accrue in the active arm of the trial. Counterfactual placebo HIV incidence varies. Performance is shown for working model and likelihood-based approaches, assuming log link function for marginal incidences. The conditional correlation ρ_m are distributed uniformly in $(0.4, 0.5)$, while the estimation procedure assumes conditional independence.

			$\rho = 0.98$			$\rho = 0.5$		
HIV incidence (cases per 100 person-years)			3	4.5	6	3	4.5	6
Exposure marker incidence (cases per 100 person-years)			7.1	11.8	17.0	4.8	13.2	26.7
			Working model approach					
$M = 10$	$n_x = 2000$	Bias × 100	0.00	0.01	0.03	0.11	0.12	0.37
		Standard deviation × 100	0.31	0.37	0.50	1.01	1.07	2.19
		Coverage (%)	96.4	97.1	97.1	95.4	95.2	95.5
	$n_x = 4000$	Bias × 100	0.00	0.02	0.04	0.16	0.08	0.37
		Standard deviation × 100	0.28	0.34	0.45	1.03	1.04	2.22
		Coverage (%)	95.5	96.5	96.8	95.3	95.3	95.1
$M = 20$	$n_x = 2000$	Bias × 100	0.01	0.01	0.02	0.06	0.04	0.16
		Standard deviation × 100	0.25	0.30	0.37	0.64	0.71	1.35
		Coverage (%)	95.7	96.3	96.6	94.3	95.8	95.1
	$n_x = 4000$	Bias × 100	0.00	0.01	0.02	0.05	0.06	0.19
		Standard deviation × 100	0.21	0.25	0.33	0.64	0.71	1.34
		Coverage (%)	95.1	96.2	96.8	94.7	94.9	95.2
			Likelihood-based approach					
$M = 20$	$n_x = 2000$	Bias × 100	0.03	0.05	0.07	0.11	0.14	0.26
		Standard deviation × 100	0.25	0.29	0.39	0.63	0.72	1.40
		Coverage (%)	95	95.3	93.4	94.4	93.9	93.5
	$n_x = 4000$	Bias × 100	0.03	0.04	0.07	0.11	0.13	0.24
		Standard deviation × 100	0.21	0.26	0.36	0.65	0.69	1.39
		Coverage (%)	95.7	94.7	92.7	93.7	94.1	93.9

Table S6. Bias, standard deviation, and empirical coverage for estimates of prevention efficacy (PE) based on M external cohorts used to estimate the association between HIV and an exposure biomarker with correlation ρ. A total of $n_x = 2000$ person-years follow-up accrue in the active arm of the trial. Counterfactual placebo HIV incidence and true PE vary. Performance is shown for working model and likelihood-based approaches, assuming log link function for marginal incidences. The conditional correlation ρ_m are distributed uniformly in $(0.4, 0.5)$, while the estimation procedure assumes conditional independence.

	M	PE estimate	$\rho = 0.98$			$\rho = 0.5$		
HIV incidence (cases per 100 person-years)			3	4.5	6	3	4.5	6
Exposure marker incidence (cases per 100 person-years)			7.1	11.8	17.0	4.8	13.2	26.7
			Working model approach					
$PE = 0.3$	10	Bias × 100	-0.74	-0.45	0.10	-4.04	-1.51	-2.96
$n_x = 2000$		Standard deviation × 100	13.96	11.01	9.78	28.20	19.20	26.53
		Coverage (%)	92.7	93.3	93.4	91.0	91.9	91.1
	20	Bias × 100	-0.65	-0.18	-0.12	-1.64	-0.64	-1.14
		Standard deviation × 100	13.24	10.61	9.11	19.71	14.18	17.4
		Coverage (%)	92.9	92.9	93.5	92.9	94.1	92.6
$PE = 0.6$	10	Bias × 100	-0.20	-0.10	-0.01	-1.86	-0.71	-1.62
$n_x = 2000$		Standard deviation × 100	9.73	7.80	6.87	17.32	12.14	15.83
		Coverage (%)	93.1	93.3	93.8	92.3	92.6	91.7
	20	Bias × 100	-0.45	0.05	-0.10	-0.74	-0.45	-0.69
		Standard deviation × 100	9.39	7.55	6.53	12.47	9.24	10.64
		Coverage (%)	93.1	93.0	93.6	93.6	94.0	93.6
$PE = 0.75$	10	Bias × 100	-0.03	0.06	0.04	-0.99	-0.29	-0.93
$n_x = 2000$		Standard deviation × 100	7.30	5.77	5.13	11.70	8.12	10.30
		Coverage (%)	93.1	93.7	93.8	92.5	93.9	92.7
	20	Bias × 100	-0.19	-0.08	0.20	-0.32	-0.08	-0.21
		Standard deviation × 100	7.11	5.63	4.89	8.80	6.63	7.34
		Coverage (%)	93.0	94.0	93.5	93.7	93.9	93.4
			Likelihood-based approach					
$PE = 0.3$	20	Bias × 100	0.38	0.17	0.46	-0.31	0.57	-0.57
$n_x = 2000$		Standard deviation × 100	12.27	9.83	8.82	18.44	13.71	17.69
		Coverage (%)	94.8	95.1	94.3	93.6	93.7	91.7
$PE = 0.6$	20	Bias × 100	0.22	0.11	0.27	-0.23	0.15	0.14
$n_x = 2000$		Standard deviation × 100	8.77	7.15	6.28	11.96	8.89	10.71
		Coverage (%)	95.3	95.2	94.7	94.2	94.8	92.6
$PE = 0.75$	20	Bias × 100	0.14	-0.07	0.23	-0.14	0.04	-0.06
$n_x = 2000$		Standard deviation × 100	6.77	5.51	4.77	8.49	6.54	7.25
		Coverage (%)	95.7	95.2	94.7	94.7	94.8	94.0

Table S7. A comparison of cohort-level vs. site-level inference. Bias, standard deviation, and empirical coverage for estimated counterfactual placebo HIV incidence and prevention efficacy, based on $M = 20$ external cohorts used to estimate the association between HIV and an exposure biomarker with correlation ρ. A total of $n_x = 2000$ person-years follow-up accrue in the active arm of the trial. Counterfactual placebo HIV incidence varies. Performance is shown for working model and likelihood-based approaches, assuming log link function for marginal incidences. Each of $M = 20$ cohorts have $L = 5$ sites and the inference is compared based on analysis of $M = 20$ cohorts vs. $M \times L = 100$ sites.

			$\rho = 0.98$			$\rho = 0.5$		
HIV incidence (cases per 100 person-years)			3	4.5	6	3	4.5	6
Exposure marker incidence (cases per 100 person-years)			7.1	11.8	17.0	4.8	13.2	26.7
Est. counterfactual placebo HIV incidence								
			Working model approach					
$M = 20$	Cohort-level	Bias × 100	-0.01	-0.02	-0.03	0.13	0.07	0.26
		Standard deviation × 100	0.33	0.38	0.53	1.03	1.05	2.22
		Coverage (%)	96.1	97.5	97.2	95.8	95.5	94.5
$M \times L = 100$	Site-level	Bias × 100	-0.01	-0.09	-0.18	0.02	-0.08	-0.19
		Standard deviation × 100	0.23	0.26	0.3	0.31	0.34	0.60
		Coverage (%)	94.2	94.3	94.3	94.4	94.6	93.6
			Likelihood-based approach					
$M = 20$	Cohort-level	Bias × 100	-0.01	-0.02	-0.02	0.11	0.09	0.25
		Standard deviation × 100	0.30	0.36	0.49	1.00	1.09	2.12
		Coverage (%)	95.4	96.5	96.8	95.4	95.4	95.7
$M \times L = 100$	Site-level	Bias × 100	0.04	0.06	0.07	0.19	0.24	0.30
		Standard deviation × 100	0.24	0.31	0.45	0.32	0.33	0.61
		Coverage (%)	93.0	90.6	89.6	90.3	89.4	91.8
Est. PE, PE=0.6								
			Working model approach					
$M = 20$	Cohort-level	Bias × 100	-0.46	-0.33	-0.42	-1.04	-0.44	-1.05
		Standard deviation × 100	9.49	7.47	6.56	12.63	9.45	11.1
		Coverage (%)	93.3	93.8	93.9	93.8	93.7	93.3
$M \times L = 100$	Site-level	Bias × 100	-0.38	-1.04	-1.28	-0.09	-0.82	-1.59
		Standard deviation × 100	9.38	7.59	6.52	9.42	7.72	7.38
		Coverage (%)	92.9	93.5	93.8	94.2	94.0	93.9
			Likelihood-based approach					
$M = 20$	Cohort-level	Bias × 100	-0.33	-0.21	-0.22	-1.14	-0.62	-1.19
		Standard deviation × 100	7.48	5.95	5.29	11.8	8.30	10.38
		Coverage (%)	93.3	93.9	93.6	93.0	92.9	91.9
$M \times L = 100$	Site-level	Bias × 100	0.35	0.21	0.25	2.13	2.07	1.54
		Standard deviation × 100	8.68	7.25	6.68	8.67	6.74	6.82
		Coverage (%)	95.3	94.3	93.3	94.7	95.2	93.5

Please use Adobe Reader or Adobe Acrobat to complete your agreement.

JOURNAL CONTRIBUTOR'S PUBLISHING AGREEMENT

To be completed by the owner of copyright in the Contribution

Title of Article	: Estimating Counterfactual Placebo HIV Incidence in HIV Prevention Trials Without Placebo Arms Based on Markers of HIV Exposure
Journal	: Clinical Trials
All Author(s)	: Yifan Zhu; Fei Gao; David V Glidden; Deborah Donnell; Holly Janes
Corresponding Author	: Fei Gao
Corr. Author Address	: fgao@fredhutch.org

Please read the full terms and conditions on the following pages, then complete, sign and return all pages of this Agreement to the **Journal's Editorial Office**.

The author who signs this Agreement certifies that he/she is authorised to sign on behalf of him/herself and in the case of a multi-authored Contribution, on behalf of all other authors of the Contribution. *The authors understand that they each have the option of signing and returning a separate copy of this Agreement. This Agreement may be signed and executed by e-mail (a scanned hard copy of the Agreement with your signature on it or a digital original copy with your electronic signature are equally acceptable), by traditional hard copy or by fax.*

EXCLUSIVE LICENCE TO PUBLISH

In consideration for publication in the above Journal, you hereby grant to the owner(s) (the 'Proprietor') of the Journal identified above (the Journal title subject to verification by SAGE Publishing ('SAGE')) the **exclusive** right and licence to produce, publish and make available and to further sub-license your article ('Article') and the accompanying abstract (all materials collectively referenced as the 'Contribution') prepared by you for the full legal term of copyright and any renewals thereof throughout the world in all languages and in all formats, and through any medium of communication now known or later conceived or developed.
If you or your funder wish your Contribution to be freely available online to non-subscribers immediately upon publication (gold open access), you can opt for it to be included in SAGE Choice, subject to payment of a publication fee. For further information, please visit SAGE Choice.

In the event you provide Supplemental Material to the Proprietor, you hereby grant to the Proprietor the **non-exclusive** right and licence to produce, publish and make available and to further sub-license the material, in whole or in part, for the full legal term of copyright and any renewals thereof throughout the world in all languages and in all formats, and through any medium of communication now known or later conceived or developed.

By signing this Contributor Agreement you agree both to the above provisions and to the Terms of the Agreement attached below.

Contributor Signature: _____ Date signed: 8/12/23

You represent that the Contribution is owned by you unless one of the following is checked:

- *If any author is an employee of the United States Government and prepared the Contribution as part of their official duties, please check here: ☐

  US Government Agency Name: _____

- If any author prepared the Contribution at the direction of their employer, please have a representative of your employer sign below, and please check here: ☐

  Employer Name: _____

  Authorized Signature: _____ Date signed: _____

***U.S. Government work.** If the Contribution was not prepared as part of the Contributor's official duties, it is not a U.S. Government work. If the Contribution was jointly authored, all the co-authors must have been U.S. Government employees at the time they prepared the Contribution in order for it to be a U.S. Government work; if any co-author was not a U.S. Government employee, then the Contribution is not a U.S. Government work. If the Contribution was prepared under a U.S. Government contract or grant, it is not a U.S. Government work - in such case, copyright is usually owned by the contractor or grantee.

If you are required to submit an addendum by your employer or research funding body, please make your request via email to ***contracts@sagepub.co.uk*** *indicating the name of the Journal and the title of your paper.*

Title:	Estimating Counterfactual Placebo HIV Incidence in HIV Prevention Trials Without Placebo Arms Based on Markers of HIV Exposure
Manuscript ID:	CT-22-0297.R4
Funding Information:	**Health/National Institute of Allergy and Infectious Diseases (NIH/NIAID)** × R01AI143357 R01CA152089 R56AI143418 UM1AI068635
Submitting Author:	Gao, Fei Sa Gao, Fei (proxy) *primary affiliation* Fred Hutchinson Cancer Center 1100 Fairview Ave N Seattle Washington 98109 United States
Authors & Institutions:	Zhu, Yifan proxy — Fred Hutchinson Cancer Center, Seattle, Washington, United States Gao, Fei proxy https://orcid.org/0000-0001-6797-5468✓ — Fred Hutchinson Cancer Center, Seattle, Washington, United States Glidden, David V. proxy — University of California, San Francisco - Epidemiology and Biostatistics, 550 16th Street , San Francisco, California 94158, United States Donnell, Deborah proxy — Fred Hutchinson Cancer Center, Seattle, Washington, United States Janes, Holly proxy — Fred Hutchinson Cancer Research Center ✓, Seattle, Washington, United States
Contact Author (populates the ##PROLE_AU	Gao, Fei Save Current Contact

THOR_..## e-mail tags):		Author: Gao, Fei (proxy)
Running Head:	Estimating Counterfactual Placebo HIV Incidence	
Keywords:	counterfactual placebo , HIV prevention , randomized controlled trial , rectal gonorrhea , trial design	