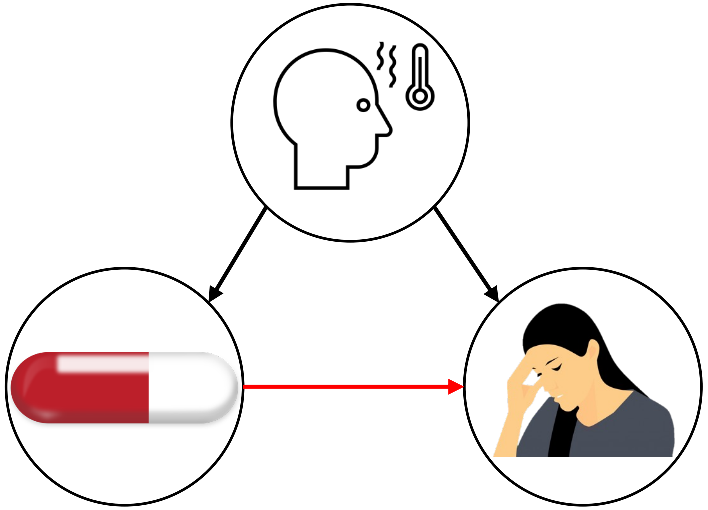
Efficient Nonparametric Estimation of Stochastic Policy Effects with Clustered Interference

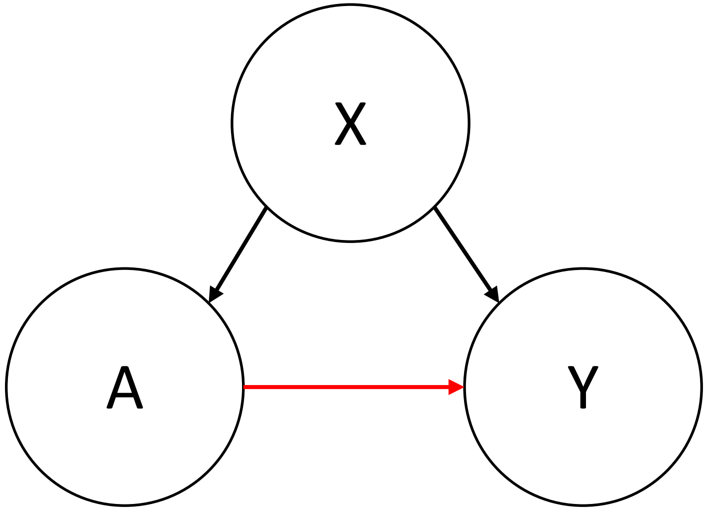
Chanhwa Lee, Donglin Zeng, Michael G. Hudgens
Department of Biostatistics, UNC

Aug 7, 2023

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Classical Causal Inference





- **Question**

  What is the difference in average potential outcomes if all individuals were treated versus not?

- **Average Treatment Effect (ATE)**

$$\text{ATE} = E[Y(1)] - E[Y(0)]$$

- **Identifiability**

$$E[Y(1)] = E[E\{Y|A=1,X\}] = E[AY/P(A=1|X)]$$

- **Estimation**
  - IPW: $\hat{E}[Y(1)] = \frac{1}{n}\sum_i \frac{A_i Y_i}{P(A_i=1|X_i)}$
  - G-formula: $\hat{E}[Y(1)] = \frac{1}{n}\sum_i E[Y_i|A_i=1,X_i]$
  - AIPW (Doubly Robust):

$$\hat{E}[Y(1)] = \frac{1}{n}\sum_i \frac{A_i\{Y_i - E[Y_i|A_i=1,X_i]\}}{P(A_i=1|X_i)} + E[Y_i|A_i=1,X_i]$$

Relax the assumption: Interference

What if the outcome is COVID19?

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-022-28825-4 OPEN

The indirect effect of mRNA-based COVID-19 vaccination on healthcare workers' unvaccinated household members

Jussipekka Salo[1,2,3,4], Milla Hägg[1,2,3,4], Mika Kortelainen[2,4,5,6], Tuija Leino[7], Tanja Saxell[2,4], Markku Siikanen[2,4] & Lauri Sääksvuori[8,9,10]✉

Mass vaccination is effective in reducing SARS-CoV-2 infections among vaccinated individuals. However, it remains unclear how effectively COVID-19 vaccines prevent people from spreading the virus to their close contacts. Using nationwide administrative datasets on SARS-CoV-2 infections, vaccination records, demographics, and unique household IDs, we conducted an observational cohort study to estimate the direct and indirect effectiveness of mRNA-based COVID-19 vaccines in reducing infections among vaccinated healthcare workers and their unvaccinated household members. Our estimates for adults imply indirect effectiveness of 39.1% (95% CI: −7.1% to 65.3%) two weeks and 39.0% (95% CI: 18.9% to 54.0%) eight weeks after the second dose. We find that the indirect effect of mRNA-based COVID-19 vaccines within households is smaller for unvaccinated children than for adults and statistically insignificant. Here, we show that mRNA-based COVID-19 vaccines are associated with a reduction in SARS-CoV-2 infections not only among vaccinated individuals but also among unvaccinated adult household members in a real-world setting.

REPORT

CORONAVIRUS

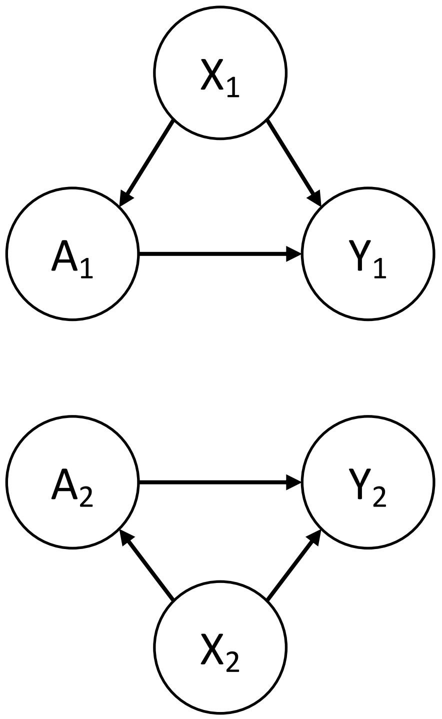
Vaccination with BNT162b2 reduces transmission of SARS-CoV-2 to household contacts in Israel

Ottavia Prunas[1,2]*, Joshua L. Warren[2,3]†, Forrest W. Crawford[2,3,4,5,6]†, Sivan Gazit[7], Tal Patalon[7], Daniel M. Weinberger[1,2]‡, Virginia E. Pitzer[1,2]‡

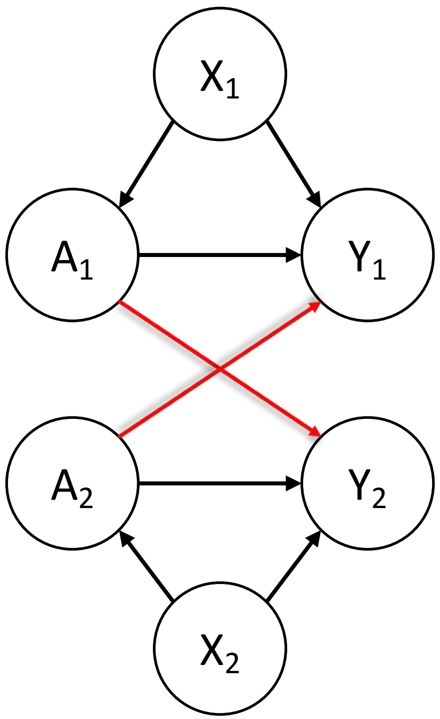
The effectiveness of vaccines against COVID-19 on the individual level is well established. However, few studies have examined vaccine effectiveness against transmission. We used a chain binomial model to estimate the effectiveness of vaccination with BNT162b2 [Pfizer-BioNTech messenger RNA (mRNA)-based vaccine] against household transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Israel before and after emergence of the B.1.617.2 (Delta) variant. Vaccination reduced susceptibility to infection by 89.4% [95% confidence interval (CI): 88.7 to 90.0%], whereas vaccine effectiveness against infectiousness given infection was 23.0% (95% CI: −11.3 to 46.7%) during days 10 to 90 after the second dose, before 1 June 2021. Total vaccine effectiveness was 91.8% (95% CI: 88.1 to 94.3%). However, vaccine effectiveness is reduced over time as a result of the combined effect of waning of immunity and emergence of the Delta variant.

1. Salo, J., Hägg, M., Kortelainen, M. *et al.* The indirect effect of mRNA-based COVID-19 vaccination on healthcare workers' unvaccinated household members. *Nat Commun* **13**, 1162 (2022).
2. Ottavia Prunas et al. ,Vaccination with BNT162b2 reduces transmission of SARS-CoV-2 to household contacts in Israel. *Science* **375**,1151-1154(2022).

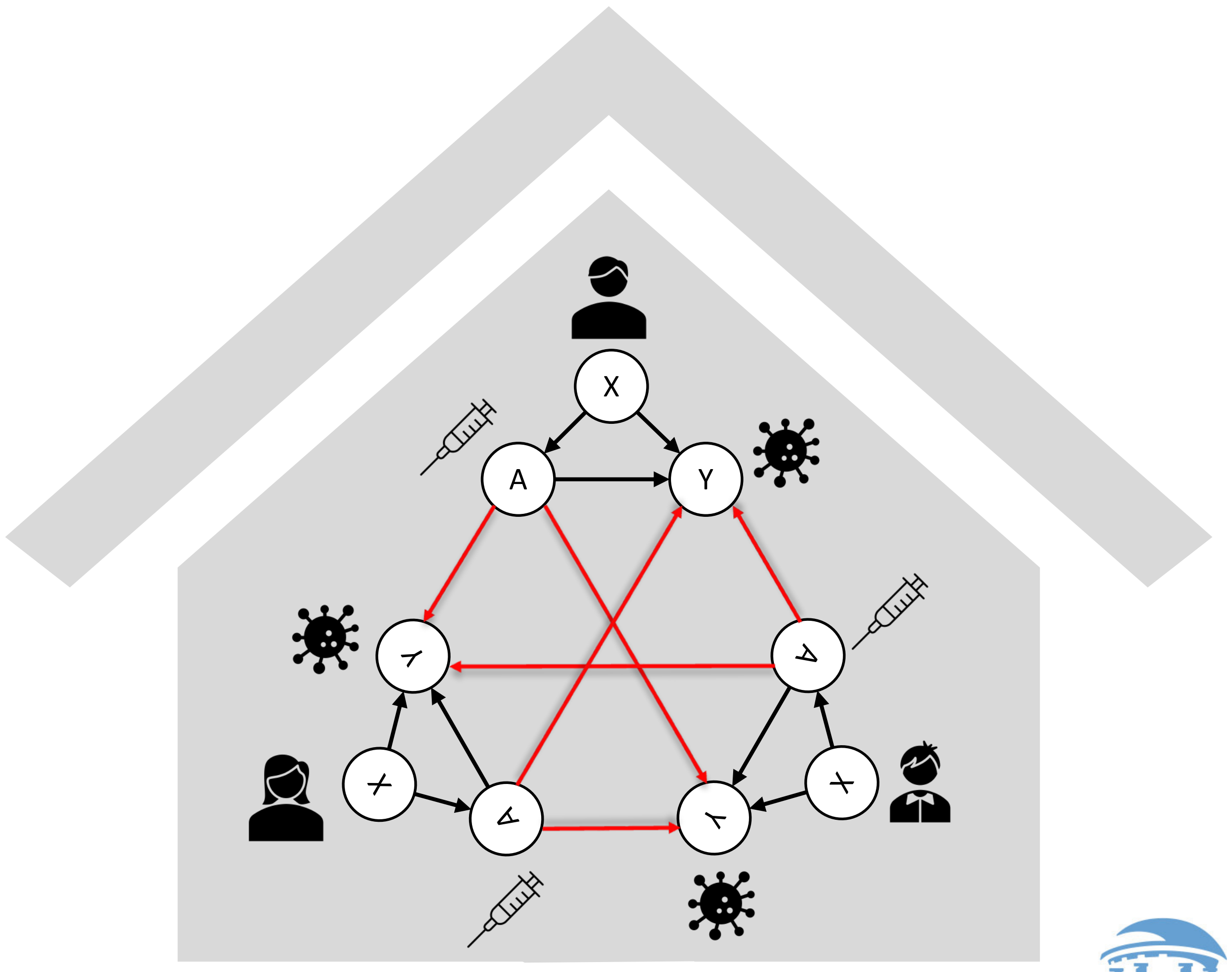
Clustered Interference



No interference



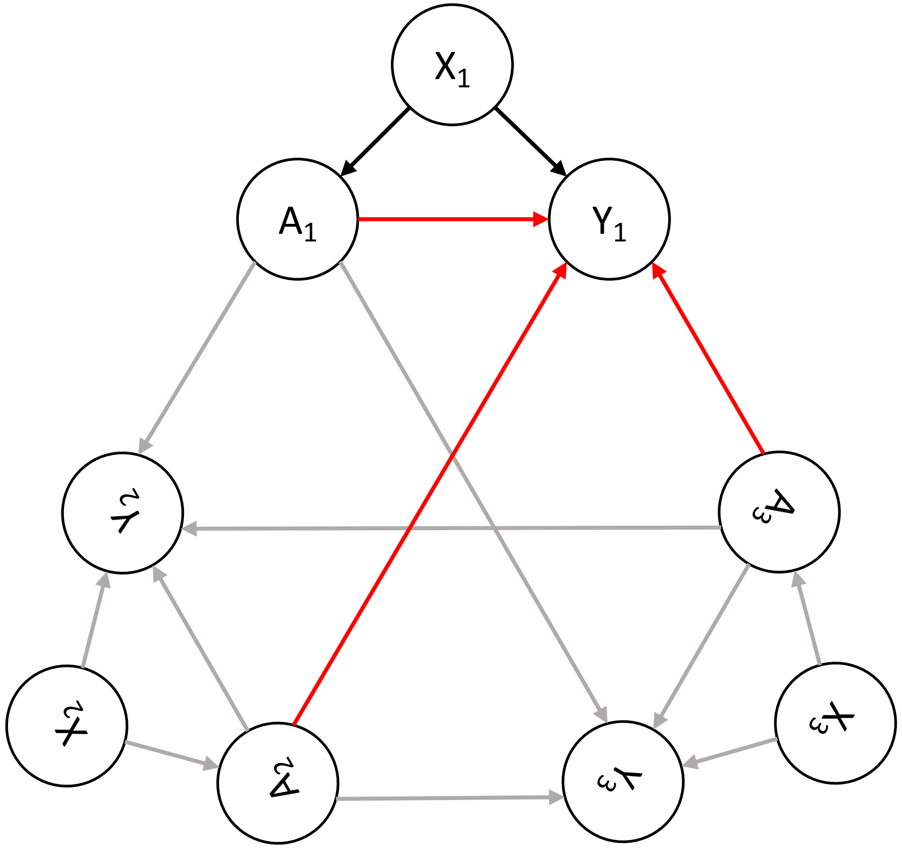
Interference



Clustered Interference

New definition for Clustered Interference

Everything is on cluster level



$Y_1(a_1, a_2, a_3)$, NOT $Y_1(a_1)$ anymore!

- **Observed data**
  - m clusters, N_i individuals in cluster $i \in \{1, \dots, m\}$
  - Unit j in cluster i,
  - $Y_{ij} \in \mathbb{R}$: outcome, $A_{ij} \in \{0,1\}$: treatment, $X_{ij} \in \mathbb{R}^p$: confounders
  - $\boldsymbol{Y_i} = (Y_{i1}, \dots, Y_{iN_i})$
  - $\mathcal{A}(N_i) = \{0,1\}^{N_i}$: set of all length N_i binary vectors

- **Potential outcome**
  - $Y_{ij}(\boldsymbol{a_i})$: potential outcome for unit j in cluster i when individuals in the cluster receives treatment (or not) according to $\boldsymbol{a_i} \in \mathcal{A}(N_i)$
  - $Y_{ij}(\boldsymbol{a_i}) = Y_{ij}(a_{ij}, \boldsymbol{a_{i(-j)}})$,
    $\boldsymbol{a_{i(-j)}} = (a_{i1}, \dots, a_{i(j-1)}, a_{i(j+1)}, \dots a_{iN_i})$

Causal Question and Causal Estimand

- **Risk of COVID19 if all were vaccinated?**

$$E\left\{\frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}(1,\ldots,1)\right\}$$

  - Not realistic!
- **Risk of COVID19 when every unit has 50% chance of getting vaccinated? i.e., 50% vaccine coverage**

$$\mu(\alpha) = E\left\{\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{\boldsymbol{a_i}\in\mathcal{A}(N_i)} Y_{ij}(\boldsymbol{a_i})Q_B(\boldsymbol{a_i};\alpha)\right\}$$

$$Q_B(\boldsymbol{a_i};\alpha) = \prod_{j=1}^{N_i}\alpha^{a_{ij}}(1-\alpha)^{1-a_{ij}}$$

  - $\mu(0.7)$ vs. $\mu(0.3)$: **Overall risk** of COVID19 when 70% vaccine coverage vs. 30% vaccine coverage
- **Risk of COVID19 if vaccinated when other units have 50% chance of getting vaccinated?**

$$\mu_1(\alpha) = E\left\{\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{\boldsymbol{a_{i(-j)}}\in\mathcal{A}(N_i-1)} Y_{ij}(1,\boldsymbol{a_{i(-j)}})Q_B(\boldsymbol{a_{i(-j)}};\alpha)\right\}$$

  - $\mu_1(0.7)$ vs. $\mu_1(0.3)$: Risk of COVID19 **if vaccinated** when 70% vaccine coverage vs. 30% vaccine coverage
- Similarly define $\mu_0(\alpha)$: **Indirect effect** from vaccinated individuals!

1. Liu, Lan, et al. "Doubly robust estimation in observational studies with partial interference." *Stat* 8.1 (2019): e214.
2. Park, Chan, and Hyunseung Kang. "Efficient semiparametric estimation of network treatment effects under partial interference." *Biometrika* 109.4 (2022): 1015-1031.

Estimation

- **Inverse probability weighting (IPW)**

$$\hat{\mu}^{IPW}(\alpha) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{Y_{ij}Q_B(\boldsymbol{A_i};\alpha)}{\hat{P}(\boldsymbol{A_i}|\boldsymbol{X_i},N_i)}$$

- **G-formula**

$$\hat{\mu}^{G}(\alpha) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{\boldsymbol{a_i}\in\mathcal{A}(N_i)}\hat{E}(Y_{ij}|\boldsymbol{A_i}=\boldsymbol{a_i},\boldsymbol{X_i},N_i)Q_B(\boldsymbol{a_i};\alpha)$$

- **Augmented IPW (Doubly Robust)**

$$\hat{\mu}^{AIPW}(\alpha) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\sum_{j=1}^{N_i}\left[\sum_{\boldsymbol{a_i}\in\mathcal{A}(N_i)}\hat{E}(Y_{ij}|\boldsymbol{A_i}=\boldsymbol{a_i},\boldsymbol{X_i},N_i)Q_B(\boldsymbol{a_i};\alpha)+\frac{Q_B(\boldsymbol{A_i};\alpha)}{\hat{P}(\boldsymbol{A_i}|\boldsymbol{X_i},N_i)}\left\{Y_{ij}-\hat{E}(Y_{ij}|\boldsymbol{A_i},\boldsymbol{X_i},N_i)\right\}\right]$$

- Need to estimate nuisance functions $P(\boldsymbol{A_i}|\boldsymbol{X_i},N_i)$ (cluster-level propensity score) and $E(Y_{ij}|\boldsymbol{A_i},\boldsymbol{X_i},N_i)$ (cluster-level outcome regression) if unknown
  - Parametric: GLM (if Y_{ij}'s (or A_{ij}'s) are independent). Otherwise, mixed effect model
  - **Nonparametric & ML**: Mixed effect ML, Smoothed kernel regression, Super Learner

Revisit Causal Question

- **Every unit has 50% chance of getting vaccinated**
  - Better to treat at-risk units more!
  - Allow for propensity depends on covariates
  - Shift observed propensity score
- **Risk of COVID19 when the odds of vaccination were 2 times the observed odds?**
  - Incremental Propensity Score Interventions (Kennedy 2019) extension to Clustered Interference setting
  - Propensity score of unit j in cluster i: $\pi_{ij} = P\left(A_{ij} = 1 \middle| \boldsymbol{X_i}, N_i\right)$
  - Shifted (counterfactual) propensity score: $\pi_{ij,\delta}$ from $\frac{\pi_{ij,\delta}}{1-\pi_{ij,\delta}} = \delta \times \frac{\pi_{ij}}{1-\pi_{ij}}$

$$\mu_{CIPS}(\delta) = E\left\{\frac{1}{N_i}\sum_{j=1}^{N_i} \sum_{\boldsymbol{a_i} \in \mathcal{A}(N_i)} Y_{ij}(\boldsymbol{a_i}) Q_{CIPS}(\boldsymbol{a_i} | \boldsymbol{X_i}, N_i; \delta)\right\}$$

$$Q_{CIPS}(\boldsymbol{a_i} | \boldsymbol{X_i}, N_i; \delta) = \prod_{j=1}^{N_i} \pi_{ij,\delta}^{a_{ij}} (1 - \pi_{ij,\delta})^{1-a_{ij}}$$

- $Q(\cdot | \boldsymbol{X_i}, N_i; \theta)$: probability distribution on $\mathcal{A}(N_i) = \{0,1\}^{N_i}$
  - Treatment allocation program (stochastic policy)

1. Kennedy, Edward H. "Nonparametric causal effects based on incremental propensity score interventions." Journal of the American Statistical Association 114.526 (2019): 645-656.

Estimation

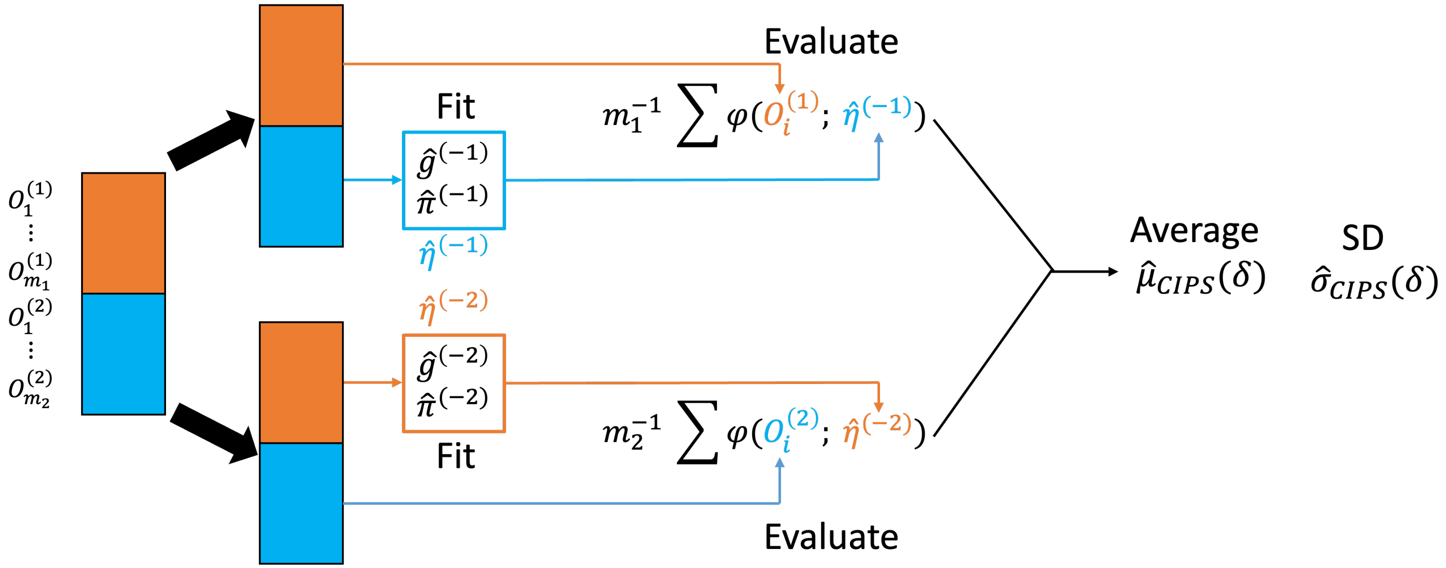
1. Find efficient influence function $\varphi^*(O_i, \eta)$

$$\frac{1}{N_i}\sum_{j=1}^{N_i}\left[\sum_{\boldsymbol{a_i}\in\mathcal{A}(N_i)} E\left(Y_{ij}\middle|\boldsymbol{A_i}=\boldsymbol{a_i},\boldsymbol{X_i},N_i\right)\{Q_{CIPS}(\boldsymbol{a_i}|\boldsymbol{X_i},N_i;\delta)+\phi_{CIPS}(\boldsymbol{A_i},\boldsymbol{X_i},N_i;\boldsymbol{a_i})\}+\frac{Q_{CIPS}(\boldsymbol{A_i}|\boldsymbol{X_i},N_i;\delta)}{P(\boldsymbol{A_i}|\boldsymbol{X_i},N_i)}\{Y_{ij}-E\left(Y_{ij}\middle|\boldsymbol{A_i},\boldsymbol{X_i},N_i\right)\}\right]-\mu_{CIPS}(\delta)$$

2. Decide nuisance functions estimators $\eta = (g, \pi)$

$$g(j,\boldsymbol{a_i},\boldsymbol{x_i},n_i) = E\left(Y_{ij}\middle|\boldsymbol{A_i}=\boldsymbol{a_i},\boldsymbol{X_i}=\boldsymbol{x_i},N_i=n_i\right),\ \pi(j,\boldsymbol{x_i},n_i)=P\left(A_{ij}=1\middle|\boldsymbol{X_i}=\boldsymbol{x_i},N_i=n_i\right)\ \Longrightarrow Q_{CIPS},\phi_{CIPS}$$

3. Implement sample splitting estimator (Chernozhukov et al. 2018)



1. Chernozhukov, Victor, et al. "Double/debiased machine learning for treatment and structural parameters." (2018): C1-C68.

Theoretical Results

Theorem

Let $\|f\| = \{\int f(\mathbf{o})^2 d\mathbb{P}(\mathbf{o})\}^{1/2}$ *denote the squared* $L_2(\mathbb{P})$ *norm. Also, let* $\delta \in \mathcal{D} = [\delta_l, \delta_u]$, *where* $0 < \delta_l < \delta_u < \infty$. *Assume the following:*

(B1) $\widehat{\pi}^{(-k)}(j, \mathbf{X}_i, N_i) \in (c, 1-c)$, $|\widehat{g}^{(-k)}(j, \mathbf{A}_i, \mathbf{X}_i, N_i)| \leq C < \infty$

(B2) $\left\| \sum_{j=1}^{N} |(\widehat{\pi}^{(-k)} - \pi)(j, \mathbf{X}, N)| \right\| = o_{\mathbb{P}}(m^{-p})$, $p \geq 0$.

(B3) $\left\| \sum_{\mathbf{a} \in \mathcal{A}(N)} \sum_{j=1}^{N} |(\widehat{g}^{(-k)} - g)(j, \mathbf{a}, \mathbf{X}, N)| \right\| = O_{\mathbb{P}}(m^{-q})$, $q \geq 0$

(B4) $\mathbb{E}(Y_{ij}^4 | \mathbf{A}_i, \mathbf{X}_i, N_i) \leq C^* < \infty$

Then,

1. *Under (B1) and (B2),* $\widehat{\mu}(\delta) = \mu(\delta) + o_{\mathbb{P}}(m^{-p})$.
2. *Under (B1)* $\sim$ *(B3) with* $p, q \geq 1/4$, $\sqrt{m}\{\widehat{\mu}(\delta) - \mu(\delta)\} \rightsquigarrow \mathbb{G}(\delta)$ in $\ell^\infty(\mathcal{D})$.
   Here, $\mathbb{G}(\cdot)$: *mean 0 GP with covariance* $\mathbb{E}\{\mathbb{G}(\delta_1)\mathbb{G}(\delta_2)\} = \mathbb{E}\{\varphi^*_{\mu(\delta_1)}\varphi^*_{\mu(\delta_2)}\}$
3. *Under (B1)* $\sim$ *(B4) with* $p, q \geq 1/4$, $\widehat{\sigma}^2(\delta) \xrightarrow{p} \sigma^2(\delta)$.
   Therefore, $\sqrt{m}\{\widehat{\mu}(\delta) - \mu(\delta)\}/\widehat{\sigma}(\delta) \xrightarrow{d} N(0,1)$.

Simulations

- $D = 1000$ simulations, each consisted of $m = 500$ clusters
- $N_i \overset{iid}{\sim} Unif\{5, 6, \ldots, 20\}, i = 1, \ldots, m$
- $C_i \sim \mathrm{N}(0, 1)$: one cluster-level covariate
- $X_{ij1} \sim \mathrm{N}(0, 1)$, $X_{ij2} \sim \mathrm{Bernoulli}(0.5)$: individual-level covariates
- $A_{ij} \sim \mathrm{Bernoulli}(p_{ij}^A)$: treatment status, $Y_{ij} \sim \mathrm{Bernoulli}(p_{ij}^Y)$: outcome

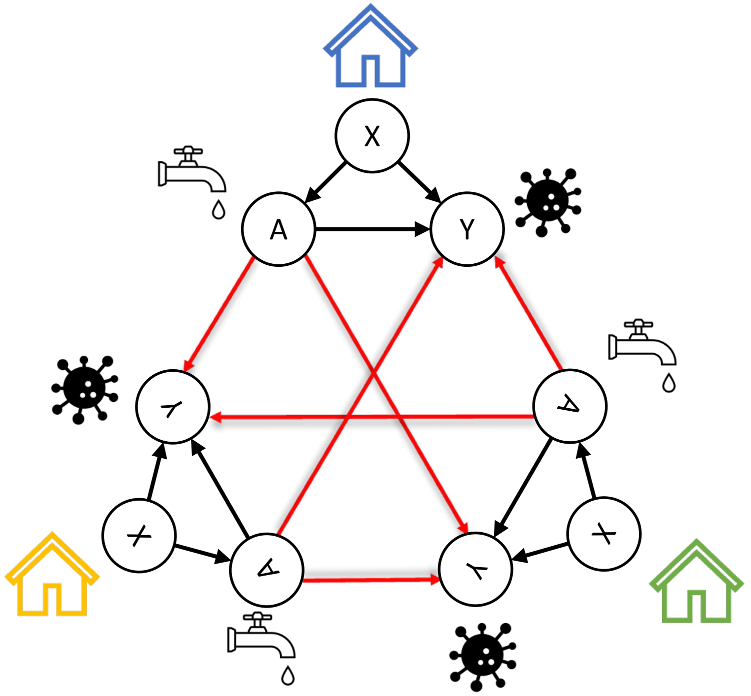
  $p_{ij}^A = \mathrm{expit}(0.1 + 0.2|X_{ij1}| + 0.2|X_{ij1}|X_{ij2} + 0.1\mathbb{1}(C_i > 0))$

  $p_{ij}^Y = \mathrm{expit}(3 - 2A_{ij} - \overline{\mathbf{A}}_{i(-j)} - 1.5|X_{ij1}| + 2X_{ij2} - 3|X_{ij1}|X_{ij2} - 2\mathbb{1}(C_i > 0))$

		Nonparametric					Parametric				
Estimand	Truth	Bias	RMSE	ASE	ESE	Cov %	Bias	RMSE	ASE	ESE	Cov %
$\mu(2)$	0.300	-0.003	0.013	0.013	0.013	94.9%	-0.010	0.017	0.013	0.014	87.1%
$\mu_1(2)$	0.224	-0.004	0.014	0.014	0.014	93.7%	-0.017	0.022	0.014	0.015	75.9%
$\mu_0(2)$	0.507	0.003	0.018	0.017	0.017	94.5%	0.025	0.031	0.019	0.019	75.6%
$DE(2)$	-0.283	-0.007	0.019	0.019	0.018	94.2%	-0.042	0.046	0.021	0.020	45.6%
$SE_1(2,1)$	-0.018	-0.002	0.010	0.010	0.010	94.3%	-0.004	0.012	0.011	0.012	93.4%
$SE_0(2,1)$	-0.022	-0.002	0.012	0.012	0.012	94.4%	-0.006	0.015	0.014	0.014	92.5%
$OE(2,1)$	-0.063	-0.003	0.009	0.009	0.009	93.9%	-0.010	0.015	0.010	0.010	77.7%
$TE(2,1)$	-0.306	-0.009	0.017	0.015	0.015	92.1%	-0.047	0.050	0.017	0.017	22.7%

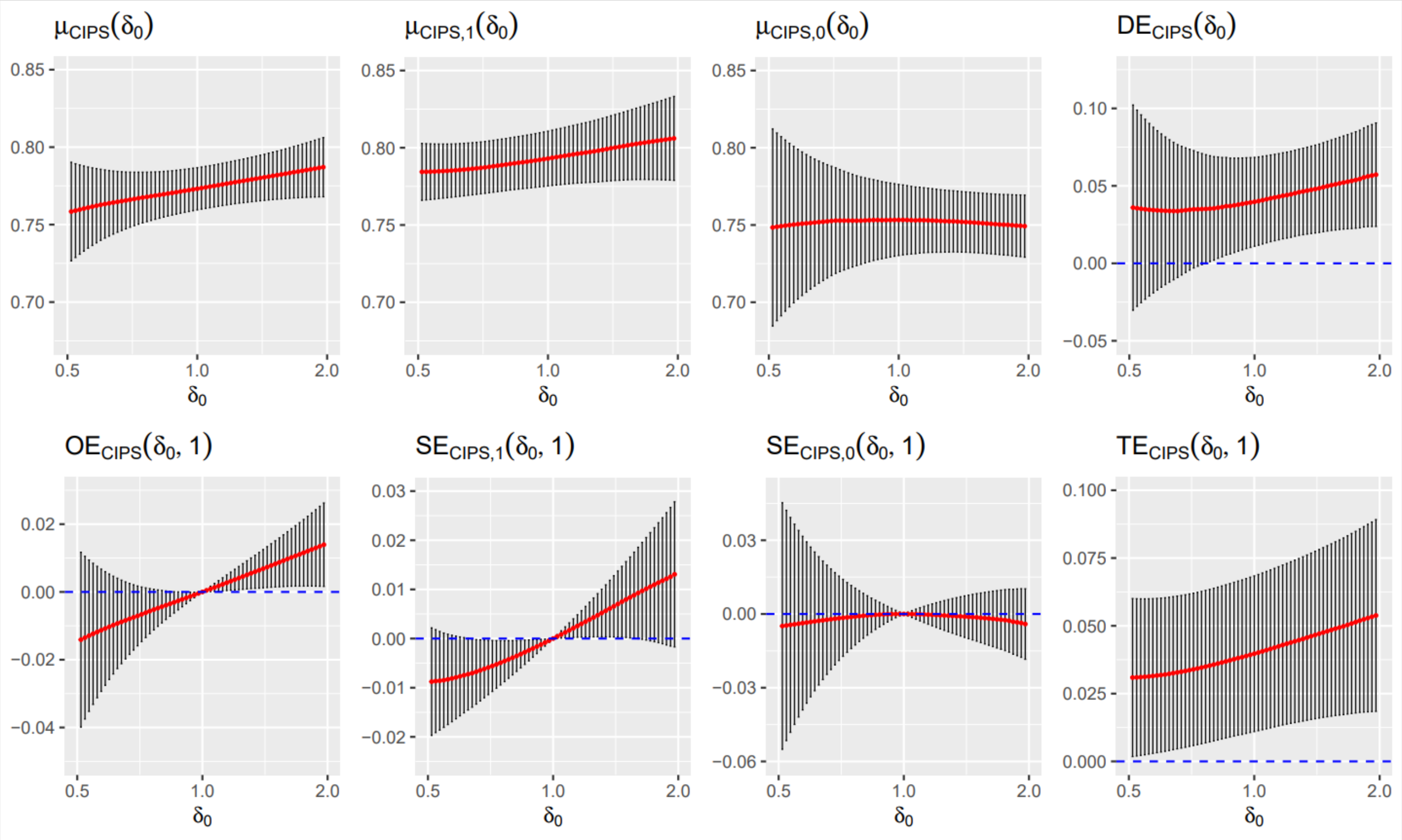
Application to Senegal Demographic Health Survey Data

- **Causal Question**
  - Whether water, sanitation, and hygiene (WASH) facilities decrease diarrhea incidence among children under clustered interference?
  - How does the diarrhea incidence change if the odds of having WASH facility change?
- **Data**
  - Cluster: Census block ($m = 1074$)
  - Unit: Household ($N_i = 2, \ldots, 12$)
  - Outcome: All children diarrhea-free
  - Treatment: WASH facility
  - Confounders: Demographic, Socioeconomic status
- **Estimation method**
  - Sample splitting estimator with Super Learner nuisance functions estimator
  - Ensemble of penalized logistic regression, spline regression, GAM, GBM, RF, neural network



1. Park, Chan, et al. "Optimal Allocation of Water and Sanitation Facilities To Prevent Communicable Diarrheal Diseases in Senegal Under Partial Interference." arXiv preprint arXiv:2111.09932 (2021).

Application to Senegal Demographic Health Survey Data



- WASH facilities prevent child diarrhea (increasing $\mu(\delta)$)
- Protective effects increase when neighboring households also have WASH facilities
- However, children from non-WASH households do not benefit from such spillover effects

Discussion

- **Future step**
  - Nonbinary treatment (continuous)
  - Censored / missing outcome (survival analysis)
  - Embedding network structure in clusters
- **Details**

Lee, Chanhwa, Donglin Zeng, and Michael G. Hudgens. "Efficient Nonparametric Estimation of Incremental Propensity Score Effects with Clustered Interference." *arXiv preprint arXiv:2212.10959* (2022).

- **References**

- Salo, J., Hägg, M., Kortelainen, M. *et al.* The indirect effect of mRNA-based COVID-19 vaccination on healthcare workers' unvaccinated household members. *Nat Commun* **13**, 1162 (2022).
- Ottavia Prunas et al. ,Vaccination with BNT162b2 reduces transmission of SARS-CoV-2 to household contacts in Israel. *Science* **375**,1151-1154(2022).
- Liu, Lan, et al. "Doubly robust estimation in observational studies with partial interference." *Stat* 8.1 (2019): e214.
- Park, Chan, and Hyunseung Kang. "Efficient semiparametric estimation of network treatment effects under partial interference." *Biometrika* 109.4 (2022): 1015-1031.
- Kennedy, Edward H. "Nonparametric causal effects based on incremental propensity score interventions." Journal of the American Statistical Association 114.526 (2019): 645-656.
- Chernozhukov, Victor, et al. "Double/debiased machine learning for treatment and structural parameters." (2018): C1-C68.
- Park, Chan, et al. "Optimal Allocation of Water and Sanitation Facilities To Prevent Communicable Diarrheal Diseases in Senegal Under Partial Interference." arXiv preprint arXiv:2111.09932 (2021).