Research Statement

Chanhwa Lee
October 28, 2024
chanhwa@email.unc.edu

1 Introduction: Causal Inference under Interference

Causal inference aims to determine whether the association between two variables—typically a treatment (A) and a response (Y)—is truly causal. Traditional methods assume no *interference*, meaning one unit's treatment does not affect others' outcomes [6]. This assumption, however, is often unrealistic. For example, a family member receiving the COVID-19 vaccine may protect not only themselves but also other family members from severe illness or death due to SARS-CoV-2 infection [18].

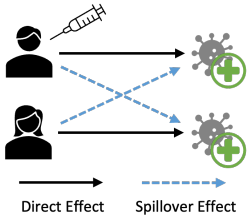

Figure 1: A visual representation of network treatment effects. A vaccinated unit is not infected due to the direct effect (black arrow). An unvaccinated unit is also not infected because of the spillover effect (blue arrow) from the vaccinated unit.

When interference is present, conventional causal inference methods cannot be directly applied. In the binary treatment case $A \in \{0,1\}$, instead of just two potential outcomes $Y(0)$ and $Y(1)$, there can be up to 2^N potential outcomes $\{Y(\mathbf{a}) : \mathbf{a} \in \{0,1\}^N\}$, where N is the number of units in a network. A unit may be affected either through direct treatment or through spillover effects from peers (Figure 1). Interference is common in the real-world settings where units interact, and ignoring it can result in misleading conclusions.

Recently, research on causal inference under interference has gained significant attention. A key example is *clustered* (or *partial*) *interference*, where units are grouped into clusters, with interference occurring only within the same cluster, but not across different clusters [8]. Clusters may be defined by spatial or temporal proximity, such as households [17] or villages [10] (see Figure 2). During my doctoral studies, I developed semiparametric efficient methods to address challenges in clustered interference, with key contributions summarized below.

2 Key Contributions

2.1 How Can We Efficiently Estimate Network Treatment Effects Across Diverse Counterfactual Scenarios?

Various causal estimands and inferential methods have been proposed to assess treatment effects under clustered interference, typically involving contrasts in average potential outcomes across different counterfactual scenarios. For instance, we may compare COVID-19 prevalence when 70% of citizens are vaccinated versus 30%, or evaluate an individual's risk based on their vaccination status while 50% of others are vaccinated. While some works [14, 17, 22] consider network treatment effects under a simple policy (Type B policy), where units independently select treatment with the same probability, this policy lacks real-world relevance because it overlooks potential heterogeneity in treatment propensities across units. Alternatively, [3, 16] propose estimands based on shifting propensity score distributions under an assumed parametric model, but model mis-specification can lead to ambiguous interpretation.

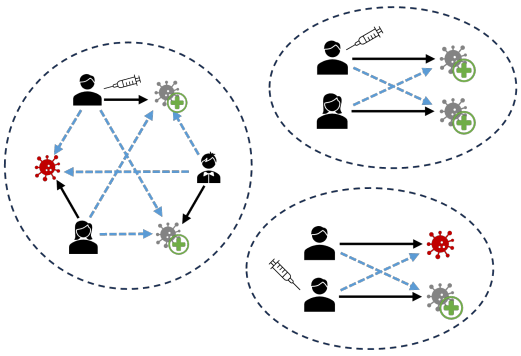

Figure 2: A visual representation of clustered interference where interfering units are divided into clusters (dotted circle).

In my work [13], published in *Journal of the American Statistical Association*, we propose nonparametric efficient estimation of network treatment effects applicable to any policy—not limited to Type B—without relying on parametric models. Drawing from [9, 23], we derive the efficient influence function (EIF) of the network treatment effect $\Psi(\mathbf{w})$, given by $\varphi(\mathbf{O}_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\varphi_{ij}(\mathbf{O}_i)$, where

$$\varphi_{ij}(\mathbf{O}_i) = \sum_{\mathbf{a}_i \in \{0,1\}^{N_i}} \{w_j(\mathbf{a}_i) + \phi_j(\mathbf{a}_i)\}\mathbb{E}(Y_{ij}|\mathbf{a}_i, \mathbf{X}_i, N_i) + \frac{w_j(\mathbf{A}_i)\{Y_{ij} - \mathbb{E}(Y_{ij}|\mathbf{A}_i, \mathbf{X}_i, N_i)\}}{\mathbb{P}(\mathbf{A}_i|\mathbf{X}_i, N_i)} - \Psi(\mathbf{w}). \quad (1)$$

Here, i indexes clusters, N_i is the size cluster i, and $O_{ij} = (Y_{ij}, A_{ij}, \mathbf{X}_{ij})$ represent the outcome, treatment, and covariates of individual j. $\mathbf{O}_i = (\mathbf{Y}_i, \mathbf{A}_i, \mathbf{X}_i)$ is the vector of data for cluster i, $\mathbf{w} = (w_1, \ldots, w_{N_i})$ is the weight function defining the target network effect (e.g., either direct or spillover), and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{N_i})$ is the EIF of the weight function $\mathbf{w}$ for policy distribution estimation.

We further propose constructing an estimator based on this EIF, using flexible, data-adaptive regression methods to avoid model mis-specification. The proposed estimator is consistent, asymptotically normal, multiply robust, and achieves the nonparametric efficiency bound. Applying this method to the Senegal Demographic

and Health Survey data [1], we find that access to private water sources or flushable toilets reduces the risk of diarrhea in children and provides an additional protective spillover effect to neighboring households.

2.2 What if Individual-Level Data are Not Available? Utilizing Cluster-Level Summary Data for Inference on Network Effects

Previous work on estimating network effects under clustered interference often assumes that all individual-level data are available for estimation. However, in some cases, individual-level data may not be collected due to budget constraints or privacy concerns, leaving only cluster-level summary data (e.g., average outcomes or mean treatment proportions). Moreover, many studies rely on inverse probability weighting (IPW) using estimated cluster-level propensity scores [3, 16, 22], which are typically computed by multiplying individual propensity scores (ranging from 0 to 1) within the same cluster. For large clusters, this product tends to become extremely small, leading to instability in IPW-based estimators.

In my work [10] (published in *Statistics in Medicine*), we propose utilizing the g-formula to account for clustered interference while relying solely on cluster-level summary data. The g-formula offers a key advantage over IPW methods by avoiding the use of potentially extreme IP weights, thereby providing greater numerical stability and bypassing positivity violations. Furthermore, under mild assumptions, network treatment effects are identifiable using cluster-level summary data, and the proposed estimator is consistent and asymptotically normal, enabling valid inference on the target effects.

Analysis of the Democratic Republic of the Congo Demographic and Health Survey data [15] using this method, which includes clusters of up to 400 individuals, suggests that increasing the proportion of children who use bed nets reduces the prevalence of malaria. Notably, we observe a protective spillover effect of bed net use on neighboring children; Figure 3 demonstrates that malaria prevalence is inversely proportional to bed net usage rate. The g-formula proved particularly effective in this analysis, as IPW methods failed to converge due to extreme weights. Additionally, the g-formula estimator relied only on cluster-level summary data—such as the number of infected children and the number of bed nets per cluster.

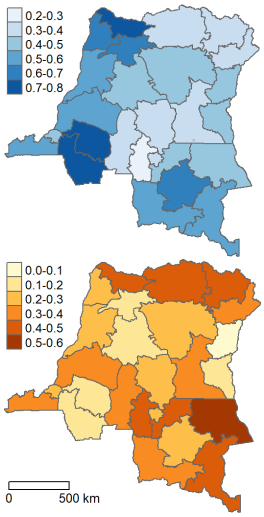

Figure 3: Province level bed net usage in the Democratic Republic of the Congo (top). Prevalence of malaria in children who do not use bed nets (bottom).

2.3 What if the Outcome is Time-to-Event? Efficiently Accounting for Confounding, Interference, and Censoring

In health science, a common interest is in estimating survival probabilities for time-to-event outcomes, such as a cholera-free rate one year after vaccination. Particularly in infectious disease contexts, methods must be developed to account for potential interference among individuals. While there is growing literature on clustered interference [3, 14, 16, 17, 22] and causal inference from observational data with censored time-to-event outcomes [7, 19, 20], only a few works combine both challenges.

In my work [12] (under review at *Journal of the Royal Statistical Society Series B: Statistical Methodology*), we propose an estimator for survival functions under various counterfactual treatment assignment scenarios that addresses both interference and right censoring. We develop an efficient estimating equation for coarsened data based on $\varphi_{ij}(\mathbf{O}_i)$ in (1), here the outcome $\mathbb{1}(T_{ij} > \tau)$ is the event-free indicator by time τ, given by

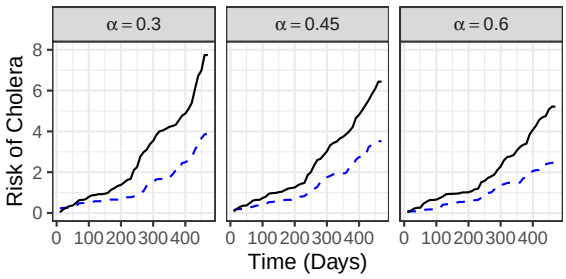

Figure 4: Estimated risk (×1000) of cholera over time when unvaccinated (black solid line) or vaccinated (blue dashed line) by vaccine coverage $\alpha \in \{0.3, 0.45, 0.6\}$. Higher the vaccine coverage, lower the risk of unvaccinated individuals.

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\sum_{j=1}^{N_i}\left[\frac{\Delta_{ij}\varphi_{ij}(\tau;\mathbf{O}_i)}{S_{ij}^{C}(Y_{ij}|\mathbf{A}_i,\mathbf{X}_i,N_i)}+\int_0^{\infty}\frac{\mathbb{E}\{\varphi_{ij}(\tau;\mathbf{O}_i)|T_{ij}\geq r,\mathbf{A}_i,\mathbf{X}_i,N_i\}}{S_{ij}^{C}(r|\mathbf{A}_i,\mathbf{X}_i,N_i)}dM_{ij}^{C}(r)\right]=0 \tag{2}$$

where m is the number of clusters, C_{ij} is the censoring time, $Y_{ij}=\min\{T_{ij},C_{ij}\}$, $\Delta_{ij}=\mathbb{1}(T_{ij}\leq C_{ij})$, S_{ij}^{C} is the survival function for the censoring time, and M_{ij}^{C} is a mean-zero martingale associated with the censoring process.

Our proposed estimator based on (2) leverages cross-fitting, allowing for nonparametric estimation of nuisance functions, while ensuring the estimator remains within the bounds of the survival function, i.e., $[0,1]$. The estimator is consistent, asymptotically normal, and multiply robust. Further, under mild conditions, it weakly converges to a Gaussian process, as do standard survival analysis estimators. The application of this

method to the Bangladesh cholera vaccine study [5] demonstrates that vaccination decreases the risk of cholera, with unvaccinated individuals experiencing a protective spillover effect from those vaccinated (See Figure 4). The direct effect of vaccination is more pronounced at lower coverage levels, while the spillover effect becomes prominent at higher coverage levels.

3 Future Directions

My previous research has centered on developing flexible, nonparametric methods under interference, and there remain many intriguing challenges within this framework. However, my interests extend beyond interference settings. In the future, I aim to explore broad and innovative topics in causal inference, particularly those that integrate nonparametric efficient estimation with machine learning techniques. Below are several research directions I propose to pursue.

3.1 Optimal Individualized Treatment Rule under Interference

Learning how to optimally assign treatments to the population is a central problem in fields such as healthcare, economics, and policy, with applications like referring patients for surgery, targeting customers with offers, or assigning students to educational programs. While there is extensive research on developing optimal individualized treatment rules (ITRs) [2, 11, 28], less attention has been given to cases where interference is present, which can lead to suboptimal or detrimental outcomes when ignored.

Building on recent advancements [27], I aim to estimate the optimal ITR under clustered interference. In particular, I will focus on maximizing survival probability by time τ under resource constraints:

$$\max_{\pi\in\Pi} \mathbb{E}\left\{\frac{1}{N_i}\sum_{j=1}^{N_i} \mathbb{1}\Big(T_{ij}\big(\{\pi(\mathbf{X}_{ij})\}_{j=1}^{N_i}\big) > \tau\Big)\right\} \text{ subject to } \mathbb{E}\left\{\frac{1}{N_i}\sum_{j=1}^{N_i}\pi(\mathbf{X}_{ij})\right\} \leq \lambda$$

where $\lambda \in [0,1]$ represents the proportion of available treatment, and $T_{ij}\big(\{\pi(\mathbf{X}_{ij})\}_{j=1}^{N_i}\big)$ is the potential survival time of unit j in cluster i if the treatment allocation follows the ITR $\pi : \mathbf{X}_{ij} \mapsto \{0,1\}$. This problem has significant real-world relevance, e.g., for maximizing cholera-free rates within a year, but with limited vaccines.

Given the complexity of modeling interference I propose an estimation method that avoids explicit modeling while accounting for potential censoring. The optimal rule will be estimated using integer programming, with theoretical guarantees on performance established via excess risk and minimax regret bounds. This framework may offer policymakers practical tools for making optimal decisions under interference, resource constraints, and censored outcomes.

3.2 Combining Causal Inference with Machine Learning

In the coming years, I plan to integrate causal inference methods with machine learning to develop robust tools for estimating causal effects in complex settings. Machine learning excels in handling high-dimensional data, nonlinearity, and intricate interactions, making it a valuable complement to traditional causal inference techniques. Recent advancements, including causal forests [25], targeted maximum likelihood estimation [24], and deep learning approaches such as adversarial learning [26] and causal representation learning [21], demonstrate promising capability to address confounding, heterogeneity in treatment effects, and high-dimensional covariates. I aim to leverage these innovations within double/debiased machine learning frameworks [4] for nonparametric nuisance parameter estimation, enabling valid inference despite complex data structures. Through this research, I seek to enhance causal inference methodologies, yielding more reliable and interpretable results across diverse fields such as healthcare and economics.

References

[1] Agence Nationale de la Statistique et de la Démographie (ANSD) and ICF. Senegal: Enquête Démographique et de Santé Continue (EDS-Continue) 2019, 2020. https://www.dhsprogram.com/pubs/pdf/FR368/FR368.pdf.

[2] S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

[3] B. G. Barkley, M. G. Hudgens, J. D. Clemens, M. Ali, and M. E. Emch. Causal inference from observational studies with clustered interference, with application to a cholera vaccine study. *The Annals of Applied Statistics*, 14(3): 1432–1448, 2020.

[4] V. Chernozhukov, D. Chetverikov, K. Kato, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C21, 2018.

[5] J. D. Clemens, J. R. Harris, D. A. Sack, J. Chakraborty, F. Ahmed, B. F. Stanton, M. U. Khan, B. A. Kay, N. Huda, M. Khan, et al. Field trial of oral cholera vaccines in bangladesh: results of one year of follow-up. *Journal of Infectious Diseases*, 158(1):60–69, 1988.

[6] D. R. Cox. *Planning of Experiments*. Wiley, New York, 1958.

[7] Y. Cui, M. R. Kosorok, E. Sverdrup, S. Wager, and R. Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211, 2023.

[8] M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[9] E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

[10] K. W. Kilpatrick, C. Lee, and M. G. Hudgens. G-formula for observational studies under stratified interference, with application to bed net use on malaria. *Statistics in Medicine*, 43(15):2852–2868, 2024.

[11] T. Kitagawa and A. Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

[12] C. Lee, D. Zeng, M. Emch, J. D. Clemens, and M. G. Hudgens. Nonparametric causal survival analysis with clustered interference. *arXiv preprint arXiv:2409.13190*, 2024.

[13] C. Lee, D. Zeng, and M. G. Hudgens. Efficient nonparametric estimation of stochastic policy effects with clustered interference. *Journal of the American Statistical Association*, (accepted), 2024. https://doi.org/10.1080/01621459.2024.2340789.

[14] L. Liu, M. G. Hudgens, B. Saul, J. D. Clemens, M. Ali, and M. E. Emch. Doubly robust estimation in observational studies with partial interference. *Stat*, 8(1):e214, 2019.

[15] Ministére du Plan et Suivi de la Mise en oeuvre de la Révolution de la Modernité (MPSMRM), Ministére de la Santé Publique (MSP), and ICF International. *République Démocratique du Congo Enquête Démographique et de Santé (EDS-RDC) 2013-2014 [Dataset]. CDPR61SD, CDGE61FL.* Rockville, Maryland, 2014. Rockville, Maryland, USA: MPSMRM, MSP and ICF International [Producers], ICF [Distributer].

[16] G. Papadogeorgou, F. Mealli, and C. M. Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019.

[17] C. Park and H. Kang. Efficient semiparametric estimation of network treatment effects under partial interference. *Biometrika*, 109(4):1015–1031, 2022.

[18] O. Prunas, J. L. Warren, F. W. Crawford, S. Gazit, T. Patalon, D. M. Weinberger, and V. E. Pitzer. Vaccination with BNT162b2 reduces transmission of SARS-CoV-2 to household contacts in Israel. *Science*, 375(6585):1151–1154, 2022.

[19] J. M. Robins and D. M. Finkelstein. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

[20] A. Rotnitzky and J. Robins. Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics*, 4:2619–2625, 2005.

[21] B. Schölkopf et al. Toward causal representation learning. *Proceedings of the National Academy of Sciences*, 118 (13):e2007669118, 2021.

[22] E. J. Tchetgen Tchetgen and T. J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

[23] A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York, NY: Springer, 2006.

[24] M. J. Van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

[25] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[26] J. Yoon, D. Jin, J. Elder, et al. Advancing the estimation of treatment effects using deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5648–5657. PMLR, 2018.

[27] Y. Zhang and K. Imai. Individualized policy evaluation and learning under clustered network interference. *arXiv preprint arXiv:2311.02467*, 2023.

[28] Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.