Research Statement

Xiaoxuan Cai

xiaoxuan.cai@yale.edu
https://xiaoxuan-cai.github.io/

My work is devoted to addressing challenging statistical and methodological problems in causal inference. My dissertation research is focused on methodological and computational issues that arise in the evaluation of infectious disease interventions. This work finds application in epidemiology, public health, and social science. Below I outline three projects that comprise my dissertation research and describe some of my future research plans.

Causal identification of intervention effects under contagion

Defining and identifying causal intervention effects for transmissible infectious disease outcomes is challenging because outcomes are dependent. In particular, a treatment (for example, vaccine) given to one individual may also affect the outcomes of others, and one infection outcome may induce further infections of others through disease transmission. This contagion between outcomes creates a seemingly interdependent relationship among outcomes and violates the conventional SUTVA assumptions. Various frameworks have been proposed to solve this dependence by randomization with the help of mediation analysis, principal stratification, or restrictive partnership models, and various estimands have been defined for the direct and indirect effect of vaccines on infection outcomes. However, these methods either preclude realistic infectious disease transmission dynamics or lose their "causal" identity with potential bias induced in large clusters.

Partnership models are popular frameworks for understanding the mechanism of infectious disease in epidemiology, and pairs even serve as the randomization units for some vaccine trials, for example, randomization trials of HIV vaccines or pneumococcal vaccines on pairs of mothers-infants. We utilize a generic partnership model to articulate the causal structure of infectious disease, define biologically meaningful estimands for both susceptibility effect and infectiousness effect, and provide non-parametric identification of causal estimands using time-to-infection data or binary outcome data. The innovation of this method is that we promote "the partner's infection time" and add it into the definition of the counterfactual outcomes, which breaks the seemingly interdependent relationships among outcomes and enables the non-parametric identification by the theory of competing risks. By realistic simulations of an HIV vaccine trial and in several theorems, we show the comparison of our estimands with popular existing quantities of vaccine's effects. These comparisons identify situations when existing estimands suffer from directional bias, in which they reveal an opposite sign of the true effect, while our estimands identify the right direction. This work is on ArXiv (https://arxiv.org/abs/1912.04151).

Causal identification of intervention under contagion by hazard models in networks

Nonparametric causal identification of intervention effects in partnership models is revealing. However, many vaccine trials are randomized in clusters and evaluated at the community level, and it can also be difficult to extend the nonparametric framework to clusters with more than two individuals. Epidemiologists have proposed hazard-based dynamic models for the time to infection that rely on the specification of the infection hazard. These models characterize the time-varying dynamics and interactions among individuals, along with the disease transmission process.

In this project, we provide causal identification results in larger clusters, and propose biologically meaningful causal estimands defined as hazard ratios. Under a class of parametric or semi-parametric models, these causal estimands can be reduced into simple terms, which are invariant to observational time and the size of the cluster. This invariance property enables direct comparisons of estimates from different studies consisting of different cluster sizes or follow-up times. Simulations of an HIV vaccine trial of different cluster sizes and under different transmission dynamics compare our causal estimands to the direct and indirect effect in the two-stage cluster randomization designs, and GEE models for clustered outcomes. The comparisons illustrate that other estimands vary with the observation time and may suffer from directional bias. Still, the scale of the bias decreases as the size of the cluster gets bigger so that the difference in exposure to infection between treated individuals and untreated individuals becomes smaller.

Semi-parametric estimation of infectious disease intervention effects

Infectious disease interventions may behave differently over seasons or contact times. For example, the flu vaccine may behave differently from an HIV vaccine. Fully parametric disease transmission models make strong assumptions about the dynamics of transmission, and fully non-parametric evaluation requires an unreasonable amount of data. In this project, we propose a semi-parametric estimation solution using non-parametric baseline hazards and parametric forms for treatments and individualistic covariates. The baseline hazards are categorized into two types: one exogenous hazard from outside the cluster, and multiple endogenous hazards from exposing to infected neighbors within the cluster. The challenge for this problem is the competing risks nature of infectious disease transmission: multiple baseline hazards are entangled together and give rise to only one outcome. Additionally, different hazards start to apply at different times, since infected network neighbors can only start transmission the disease after they become contagious at different times. We apply the Martingale theory to develop a two-step iteration procedure to estimate baseline hazards and parameters of interest. The method has applications beyond infectious disease outcomes. For any problem involves two entangled baseline hazards under different timelines, this method permits estimation of the main effects along with baseline hazards. In ongoing research, we are deriving point-wise confidence intervals for baseline hazard estimates, and extending the method to the computationally challenging setting of larger clusters.

Future research plans

I am interested in continuing to study treatment effects among dependent outcomes embedded in a network and evaluation of dynamic treatments for time-to-event outcomes. In addition to my current projects, I am also interested in novel study designs to evaluate interventions among dependent outcomes. I have broader interests in embracing causal inference with popular machine learning methods to improve estimation, especially to address public health concerns.

References

[1] Xiaoxuan Cai, Wen Wei Loh, and Forrest W. Crawford. Identification of causal intervention effects under contagion. (2019) - submitted.

[2] Xiaoxuan Cai and Forrest W. Crawford. Causal identification of intervention effects under contagion by hazard models in networks. - working paper.