Research Statement

Xiaoxuan Cai

Columbia University
xc2577@cumc.columbia.edu
https://xiaoxuan-cai.github.io/

My research is devoted to addressing statistical challenges and methodological problems in causal inference, missing data, mediation analysis, and machine learning. My work covers applications in intervention evaluation and trial design for outcomes subject to interference (e.g. infectious disease), as well as the analysis of multivariate time series of digital data from mobile devices. My current research projects include i) formalizing the causal structure of infectious outcomes and evaluating infectious disease interventions in a stochastic and interactive transmission network, where outcomes and treatments are interdependent across all participants; ii) articulating the causal relationship between "exposure time series" and "outcome time series" for non-stationary multivariate time series in the presence of time-varying confounders; iii) missing data imputation in non-stationary multivariate time series; and iv) pathway decomposition and mediation analysis for non-stationary multivariate time series. This methodological work finds applications in epidemiology, public health, psychiatry, political science, and economics. Below I introduce my main projects and future plans.

Causal identification for infectious disease intervention effects

COVID-19 has been the world's most serious health threat for the last two years, responsible for over 244 million confirmed cases and 4.95 million deaths as of October 2021. Evaluations of containment and mitigation policies, as well as vaccine effectiveness, are urgently needed to inform future policy making and resource allocation. Identifying causal effects of infectious disease interventions in a complex person-to-person interacting network is difficult because an intervention (e.g., vaccine) given to one individual may affect outcomes of others, and one infection outcome may cause additional infections of others through disease transmission.

Identification in partnership settings Models of contagion in two-person partnerships are popular frameworks for formalizing the mechanism of infectious disease transmission in epidemiology. Pairs serve as units of randomization in some vaccine trials, including randomized trials of HIV vaccines or pneumococcal vaccines in mothers-infant pairs. In Cai et al. [1], we use a generic partnership setting to formalize the causal structure of infectious disease transmission, define biologically meaningful estimands for susceptibility and infectiousness effects, 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 to allow for better control of this confounding variable. This strategy breaks the seemingly interdependent relationships among outcomes and enables non-parametric identification by the theory of competing risks. Using realistic simulations of vaccine trials and in several theorems, we compare our estimands with popular existing quantities for vaccine 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.

Non-parametric/Semi-parametric identification in network settings Some vaccine trials are randomized in clusters and evaluated at the community level. Existing statistical frameworks and estimands for infectious disease intervention evaluation, including structural transmission models, mediation-based partnership models, and two-stage randomization designs, are of limited conceptual usefulness: i) parameters in structural models lack clear causal interpretation; ii) mediation-based causal effects are confined to a restricted two-person setting; and iii) other randomization-based estimands summarize cluster-level quantities without direct correspondence to biologically meaningful effects and may suffer from potential bias. In Cai and Crawford [2], we describe a unifying formalism for defining nonparametric structural causal estimands and an identification strategy for learning about the effects of infectious disease interventions on clusters of interacting individuals when infection times are observed. This strategy unifies existing estimands across study designs and provides a framework for causal identification in both randomized and observational studies, leveraging individual treatments and covariates which may affect susceptibility and infectiousness. The identification strategy extends the simple case of one partner's infection time as a mediator in the partnership setting to a group of stochastic infection times as outcomes and mediators for each other dynamically in a cluster. We define new causal estimands – contagion, susceptibility, and infectiousness effects – and compare them to widely used direct and indirect effects in the two-stage randomized trial, in which they reveal directional bias. We further extend a semi-parametric hazard framework developed by Kenah [3, 4] to facilitate statistical estimation of parameters and causal effects, incorporating exogenous hazard from outside the cluster and endogenous hazards from exposing to infected neighbors within the cluster.

Future research plans I am interested in continuing my work on causal identification for outcomes under interference embedded in a stochastic network. I plan to extend my current research on causal identification for contagious outcomes from the ideal case of knowing infection times to more realistic scenarios. In particular, because infection times are often unknown and we only know infection happens prior to a given time due to incubation or delayed confirmation from tests, left-truncation in outcomes needs to be incorporated. Following the classical SEIR-type disease transmission models, the incubation period during which subjects are infectious without symptoms, as well as recovery during which subjects are no longer infectious, are crucial to include when evaluating causal effect of infectious disease intervention. Data augmentation may be considered, leveraging information from other trials or disease transmission dynamics. Based on the topology of a disease transmission network, I am also interested in finding super-spreaders and exploring potential optimal strategies to mitigate disease transmission. Additionally, I am interested in novel trial designs for evaluating (individual- or community-level) interventions for outcomes under interference. I am also broadly interested in the use of innovative mobile technologies (e.g., cell phone mobility data) to aid in real-time learning about how and where people become infected. This improved granularity in disease transmission surveillance has the potential to advance our understanding of the epidemiological characteristics of infectious disease and transmission dynamics within a cohort, leading in a more nuanced assessment of policy and vaccine effectiveness.

Causal inference for (non-stationary) multivariate time series from digital devices

Mobile technology (e.g., mobile phones and wearable devices) allows collection of information in real time and in naturalistic settings, presenting an enormous opportunities for scientific discoveries and advances in dynamic and personalized intervention in medical and social science. Digital data streams may contain passively collected signals (e.g. telecommunication, accelerometer, and GPS data) and actively collected information (e.g. electronic momentary assessment (EMA), survey responses, EHR data), and they constitute entangled multivariate time series embedded in a complex and potentially dynamic system. My research aims to develop statistical methods that can be used to inform dynamic and personalized health decision making. Ongoing projects include: i) causal inference for the effect of time-varying exposures/interventions on continuous outcome trajectories in the presence of time-varying confounders; ii) missing data imputation for non-stationary multivariate time series; and iii) pathway decomposition and mediation analysis of exposure effects in multivariate time series.

Causal estimands of short-term and long-term effects in multivariate time series Conceptualizing and evaluating (dynamic) causal effects for (non-stationary) multivariate time series is challenging, as subjects are continuously assigned to time-varying exposures and outcomes are continuously monitored. The majority of existing estimands are concerned with describing the effect of an exposure path/history on an outcome at a specified time point, the (time-varying) contemporaneous effect of exposure on an upcoming outcome prior to the next treatment assignment, or the evolution of the effect of a continuing intervention (also known as “shocks” in economics). Researchers in psychiatry are interested in how behavioral interventions (such as exercise and socialization) can benefit patients with serious mental illnesses. A successful behavioral intervention in mental health usually entails systematic behavioral changes that take time to develop, and as a result, the benefit of intervention may be delayed or may only appear after a period of treatment. Novel causal estimands are required in order to adequately describe the causal relationship between “exposure time series” and “outcome time series”. In Cai et al. [5], we propose a collection of causal estimands to summarize exposure effects throughout time on the outcome trajectory, including contemporaneous (or short-term) effect, lagged effects, and long-term effects, and to depict their evolution over time. All of these effects are crucial for advancing our understanding of treatment effects and providing more personalized treatment at the optimal moment. We provide a causal identification strategy and an estimation strategy based on the g-formula in the presence of exposure-covariate and outcome-covariate feedbacks in non-stationary multivariate time series. Using a multi-year observational smartphone study of bipolar and schizophrenia patients, we investigate how social support, as measured by the “degree of social contacts”, affects patients' negative mood over short- and long-term time periods, and identify specific periods when patients are more reactive to the intervention.

Missing data imputation for non-stationary multivariate time series Missing data is an ubiquitous problem in almost all fields that collect data, and is especially prevalent with mobile device data. Imputation is commonly recommended to maximize the utility of valuable data and improve estimation efficiency for quantities of interest. Most of existing imputation methods are either designed for longitudinal data with limited follow-up times or for stationary time series, which may not be suitable for many social, economic and political issues due to the system's constant evolution over time. In Cai et al. [6], we propose a novel multiple imputation method based on the state space model (SSMmp) to address missing data in multivariate time series that are potentially non-stationary. We evaluate its theoretical properties and performance in extensive simulations of both stationary and non-stationary time series under different missing mechanisms, showing its advantages over other commonly used strategies for missing data. We apply the SSMmp method in the analysis of a multi-year observational smartphone study of bipolar patients – the Longitudinal Bipolar Study – to evaluate the association between social network size and psychiatric symptoms adjusting for confounding.

Pathway decomposition in the presence of multiple mediators in non-stationary multivariate time series The effect of exposures on outcomes over time in the context of entangled exposure, outcome, and covariate time series can operate via a complex network of pathways. Disentangling and quantifying each pathway from the complex network is a daunting task that adds little to our understanding of the overall mechanism. By aggregating the direct effect of exposures on outcomes, the indirect effect of exposures on outcomes carried over by outcome autocorrelation, and the remaining indirect effect of exposures on outcomes via mediating covariate time series, we demonstrate a systematic way for decomposing complex pathways induced by exposure-covariate and outcome-covariate feedback into distinct subgroups of pathways corresponding to different mechanisms. For example consider the setting in which the treatment is social interaction, the outcome is positive feelings of patients, and the mediators of interested is medication adherence or exercise. Then, social interaction may directly improve a patient's positive feelings (direct effect), which may be sustained over time (indirect effect via outcome autocorrelation), or it may indirectly improve medication adherence or exercise, which in turn increases positive feelings (indirect effect via mediators). By grouping pathways into distinct mechanisms, we avoid the burdensome task of identifying detailed structures between variables over lags in a DAG, as well as difficulties associated with causal identification. The relative importance of each mechanism suggests possible directions for future behavioral intervention design. We provide an identification strategy for the direct effect and two indirect effects of exposures on the outcome trajectory in non-stationary time series. In a multi-year observational study of patients with bipolar disorder, we examined the relative importance of the direct effect of social support on psychiatric symptoms and its indirect effect via a mediator– physical exercise. This work contributes to the development of definitions for direct and indirect effects in complex psychiatry mediation research, which routinely involves multiple mediators and repeated measures of variables.

Future research plans The statistical exploration of digital health data from mobile devices is still in its early stages, facing both challenges and opportunities for advancements in statistical methodologies, computation, and scientific discovery. I intend to pursue the following projects to continue my research into methodological issues involving multivariate time series data from digital devices: 1) extending the state space model's multiple imputation strategy to missing data also in exposures and covariates in non-stationary multivariate time series; 2) exploring causal estimation strategies for time-varying effects of exposures by incorporating matching, inverse weighting, and machine learning algorithms to alleviate the burden of parametric assumptions; and 3) combining (non-stationary) multivariate time series from multiple individuals to inform effect evaluation while allowing for heterogeneous disease progression. I am also interested in 4) analyzing time-to event outcomes (e.g. time to relapse, risk of suicide) using continuous monitoring of exposures and covariates, 5) causal discovery using machine learning algorithms and hypothesis testing, and 6) information integration of high-dimensional digital signals (possibly from multimedia and under different time resolutions). Examples include the integration of high-dimensional multimedia telecommunication, self-evaluated social interaction in person and digitally, daily routines (e.g., frequently visited locations, hours spent outside) analyzed from GPS data to summarize social interaction and the integration of MRI data and accelerometer data to summarize sleep quality. For the first time in decades, researchers now have an extremely powerful tool for delving into this massive collection of immersed and densely measured digital data from mobile devices. I anticipate an increasing number of research trials in public health, environmental science, political science, psychiatry, and other social science disciplines in the near future that will leverage digital technology. I am excited to continue exploring statistical advances and developing new algorithms for studying scientific questions with increased precision and delivering more tailored and timely interventions using this new mobile technology.

Finally, my research is highly collaborative and interdisciplinary and is strongly motivated by challenges in public and mental health. I plan to continue exploring open statistical challenges in these fields and develop new collaborations with my colleagues, external agencies, as well as students, with the goal of having a positive impact on our society and population health.

References

[1] Xiaoxuan Cai, Wen Wei Loh, and Forrest W Crawford. Identification of causal intervention effects under contagion. *Journal of Causal Inference*, 9(1):9–38, 2021.

[2] Xiaoxuan Cai and Forrest W. Crawford. Causal identification of intervention effects under contagion by hazard models in networks. *Submitted to Annals of Statistics*, 2021.

[3] Eben Kenah. Non-parametric survival analysis of infectious disease data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):277–303, 2013.

[4] Eben Kenah. Semiparametric relative-risk regression for infectious disease transmission data. *Journal of the American Statistical Association*, 110(509):313–325, 2015.

[5] Xiaoxuan Cai, Xinru Wang, Jukka-Pekka Onnela, Justin T. Baker, and Linda Valeri. Causal inference for multivariate time series data in n-of-1 studies. *Working progress*, 2021.

[6] Xiaoxuan Cai, Xinru Wang, Justin T. Baker, Jukka-Pekka Onnela, and Linda Valeri. Multiple imputation via state space model for missing data in non-stationary multi-variate time series. *Manuscript in preparation*, 2021.