Xiaoxuan Cai

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

EDUCATION

Yale University, *Ph.D. in Biostatistics* 08/2020 – 08/2015
Advisor: Forrest W. Crawford
Yale University, *M.S. in Biostatistics* 05/2015 – 08/2013
Advisor: Peter M. Aronow
Peking University, *B.S. in Biology, B.A. in Economics* 07/2013 – 09/2009

TRAINING

Postdoctoral Research Fellow, Department of Biostatistics, Columbia University **New York, NY**
Advisor: Linda Valeri 08/2020 – Present

RESEARCH EXPERIENCE

Department of Biostatistics (Advisor: Linda Valeri), Columbia University **New York, NY**
Postdoctoral Research 08/2020 – Present

Causal identification of time-varying short-term and long-term effects in non-stationary time series

- Formulated a collection of causal estimands for systematically summarizing the time-varying behavior of exposure effects on the outcome trajectory.
- Based on the state space model, adjusted dynamic exposure-covariate and outcome-covariate feedbacks via g-formula in non-stationary multivariate time series
- Demonstrated the potential for combining short- and long-term effects to inform optimal treatment strategy design.
- Illustrated its use in a multi-year observational smartphone study of bipolar patients, identifying time-varying contemporaneous and lagged effects of social support on negative moods, as well as periods during which patients are more susceptible to social support influence.

Missing data imputation for non-stationary multivariate time series

- Developed a novel iterative multiple imputation method based on the state space model (SSMmp) to address the problem of missing data in multivariate time series that are potentially non-stationary.
- Established the theoretical properties of SSMmp in terms of unbiasedness and increased efficiency compared to the complete case analysis
- Assessed SSMmp's empirical performance in simulations of stationary and non-stationary time series under multiple missing mechanisms (MCAR, MAR, and MNAR), demonstrating its superiority to other commonly used strategies for missing data.
- Using a multi-year observational smartphone survey, applied SSMmp to the examination of the association between social network size and mental symptoms in bipolar patients.

Pathway decomposition and mediation analysis in non-stationary multivariate time series

- Decomposed complex pathways involving entangled multivariate time series of exposures, outcomes, and covariates with exposure-covariate and outcome-covariate feedbacks into exclusive subgroups of pathways, 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.
- Provided causal identification for the direct effect and two indirect effects of exposures on the outcome trajectory in non-stationary time series
- Calculated the direct effect and two indirect effects of social support on negative feelings in bipolar patients using a multi-year observational smartphone study, demonstrating the relative magnitude of each mechanism.

Department of Biostatistics (Advisor: Forrest W. Crawford), Yale University **New Haven, CT**
Ph.D. Research 12/2015 – 08/2020

Causal identification for the infectious disease intervention under contagion

- Articulated the infectious disease causal structure in a generic partnership setting without restrictions on who-infects-whom, in which interventions and outcomes are all interdependent (referred to as "contagion")
- Defined biologically meaningful estimands for susceptibility and infectiousness effect, and proved how they relate to (and correct bias in) widely used estimands in epidemiology
- Provided non-parametric identification of those causal estimands using time-to-infection data or binary outcome data, with no requirement on study design or disease transmission models.
- Illustrated proposed estimands using a realistic simulation of an HIV vaccine trial and compared their performance to that of other widely used epidemiological estimands

Causal identification of infectious disease intervention effects in a clustered population

- Described a unifying formalism for defining nonparametric structural causal estimands for learning about infectious disease intervention effects in clusters of interacting individuals, which conglomerates existing estimands from varied designs.
- Provided non-parametric identification strategies for proposed estimands, when SUTVA assumptions are violated due to infectious outcomes being transmissible and interventions having an indirect effect.
- To facilitate statistical inference in finite samples, a semi-parametric class of pairwise Cox-type transmission hazard models was adopted.
- Compared proposed estimands to existing popular estimands for clustered randomized trials in epidemiology, using extensive simulations under a variety of randomized and observational vaccine trial designs.

Takeda Pharmaceutical Company **Cambridge, MA**
Summer Intern to the Statistics Department 06/2016 – 08/2016

- Implemented methods for mining Twitter data (~300,000 tweets) to identify the most influential users for a given topic
- Implemented topic classification algorithm for extracting main topics from a collection of raw Twitter messages; Adapted the method to use the Twitter API to perform real-time analysis
- Led discussions on text mining and information extraction techniques in weekly department meetings
- Presented summer project to the Takeda Statistics Department with an audience of over 100 people

School of Medicine, (Dr. Kimberly Yonkers Lab), Yale University **New Haven, CT**
Graduate Student Researcher 03/2014 – 04/2015

Identify risk factors for cocaine addiction among pregnant women

- Determined model most appropriate for analysis based on assumptions inherent of each approach:
  - Mixed-effects regression model for ordinal outcomes
  - Generalized estimation equation (GEE) model for ordinal outcomes and GEE model for negative binomial distribution
  - Zero-inflated and zero-altered Poisson model for correlated data
- For each model, performed single variable exploration and backward selection.
- Identified age of first use, social support, baseline cocaine use, marijuana usage as key potential risk factors

Department of Political Science (Dr. Peter M. Aronow Lab), Yale University **New Haven, CT**
Graduate Student Researcher 10/2014 – 05/2015

- Applied LASSO variable selection method to Generalized Estimation Equation (GEE) model for binary response variable and write R codes for implementation.
- Applied to longitudinal dataset about cocaine addiction during pregnancy, and compare analysis results with the GEE model for negative binomial distribution both in real data and simulations

PAPERS

Xiaoxuan Cai, Jukka-Pekka Onnela, Justin T. Baker, Linda Valeri (2021). Causal inference for multivariate time series data in N-of-1 studies. *(working progress)*

Xiaoxuan Cai, Xinru Wang, Justin T. Baker, Jukka-Pekka Onnela, Linda Valeri (2021) State space model multiple imputation for missing data in non-stationary multivariate time series. *(Manuscript accepted by NeurIPS 2021 Workshop on Causal Inference Challenges in Sequential Decision Making: Bridging Theory and Practice)*

Xiaoxuan Cai, Eben Kenah, and Forrest W. Crawford (2020) Causal identification of infectious disease intervention effects in a clustered population. *Submitted to Annuals of Statistics*

Xiaoxuan Cai, Wen Wei Loh, and Forrest W. Crawford (2021) Identification of causal intervention effects under contagion. *Journal of Causal Inference.*

Regina Melendez, **Xiaoxuan Cai,** Cristine Hine, et al. (2015) Correlates of Cocaine Use in Pregnancy. *Yale Medicine Thesis Digital Library.*

PRESENTATIONS

2022 Invited session, 2022 Joint Statistical Meetings, Washington, DC
2021 Invited session, NeurIPS 2021 Workshop on Causal Inference Challenges in Sequential Decision Making: Bridging Theory and Practice.
2021 E-poster session, 3rd Annual Health Data Science Symposium, Boston, MA
2021 Presentation, 2021 Joint Statistical Meetings, Virtual
2021 Invited session, 42th International Society of Clinical Biostatistics, Lyon, France
2021 Invited session, 2021 ENAR, Virtual
2020 Causal inference learning group, School of Public Health, Columbia University, Virtual
2020 Biostatistics department seminar, School of Public Health, Columbia University, Virtual
2020 Invited session, 2020 Women in Statistics and Data Science Conference, Virtual
2020 Invited session, 2020 Joint Statistical Meetings, Philadelphia, PA
2019 Research presentation, School of Public Health, Yale University, New Haven, CT
2019 Contributed session, 2019 Joint Statistical Meetings, Denver, CO
2019 Invited session, 33nd New England Statistics Symposium, Hartford, CT
2019 Invited session, 2019 ENAR, Philadelphia, PA
2018 Invited session, 2018 Women in Statistics and Data Science Conference, Cincinnati, OH
2018 Poster session, 32nd New England Statistics Symposium, Amherst, MA
2018 Speed session, 2018 Joint Statistical Meetings, Vancouver, BC

AWARDS AND RECOGNITION

2019 Young Investigator Award, ASA Section on Statistics in Epidemiology
2018 Travel Award, Women in Statistics and Data Science conference
2018 Travel Award, Summer Institutes at the University of Washington
2018 MassMutual Student Poster Award, 32nd New England Statistics Symposium
2015 Fellowship from Takeda pharmaceutical Company
2013 Huirong Li Scholarship, Peking University

MENTORSHIP AND DEPARTMENT SERVICE

Co-mentored one master student (Xinru Wang) with Dr. Valeri, Columbia University 03/2021 – Present
Co-organized the Causal Inference Learning Group with Dr. Valeri, Columbia University 08/2020 – Present

LEADERSHIP and TEACHING EXPERIENCE

LEADERSHIP

Chairperson of the Connecticut Peking University Alumni 05/2017 – 09/2018

Class President, Department of Life Sciences, Peking University 09/2010 – 06/2012

Minister of the Academic Department, the Student Union, Peking University 09/2010 – 07/2011

Member of the School's Debate Team, Peking University 09/2009 – 07/2010

TEACHING EXPERIENCE

Yale University, Department of Biostatistics **New Haven, CT**

Teaching Assistant (Course: Applied Survival Analysis, Intro to Stat Thinking, Applied Regression Analysis, Biostatistics in Clinical Investigation) 08/2014 – 05/2018

Peking University, Department of Biology **Beijing, China**

Teaching Assistant (Course: Biostatistics) 09/2012 – 01/2013

SKILLS AND TRAININGS

Additional Trainings: Causal Inference and Big Data Summer Camp, University of Pennsylvania (2017)
Summer Institute in Statistics and Modeling in infectious disease, University of Washington (2018)

Skills: Python, R, C, C++, SPSS, Stata, Matlab, Git, LaTeX;

Language: Chinese Mandarin (native), English (Fluent)