Xihong Lin, is Professor and former Chair of the Department of Biostatistics, Coordinating Director of the Program in Quantitative Genomics at the Harvard T. H. Chan School of Public Health, and Professor of the Department of Statistics at the Faculty of Arts and Sciences of Harvard University, and Associate Member of the Broad Institute of Harvard and MIT.

Dr. Lin is an elected member of the National Academy of Medicine. She received the 2002 Mortimer Spiegelman Award from the American Public Health Association, and the 2006 Committee of Presidents of Statistical Societies (COPSS) Presidents’ Award and the 2017 COPSS FN David Award. She is an elected fellow of American Statistical Association (ASA), Institute of Mathematical Statistics, and International Statistical Institute.

Dr. Lin’s research interests lie in development and application of statistical and computational methods for analysis of massive data from genome, exposome and phenome, and scalable statistical inference and learning for big genomic, epidemiological and health data.

**Title:** Regression Models for Understanding COVID-19 Epidemic Dynamics with Integrated Data

**Abstract:** Modeling infectious disease dynamics has been critical throughout the COVID-19 pandemic. Of particular interest are the incidence, prevalence, and effective reproductive number (Rt). Estimating these quantities is challenging due to under-ascertainment, unreliable reporting, and time lags between infection, onset, and testing. We propose a Multilevel Epidemic Regression Model to Account for Incomplete Data (MERMAID) to jointly estimate Rt, ascertainment rates, incidence, and prevalence over time in one or multiple regions. Specifically, MERMAID allows for a flexible regression model of Rt that can incorporate geographic and time-varying covariates. To account for under-ascertainment, we (a) model the ascertainment probability over time as a function of testing metrics and (b) jointly model data on confirmed infections and population-based serological surveys. To account for delays between infection, onset, and reporting, we model stochastic lag times as missing data, and develop an EM algorithm to estimate the model parameters. We evaluate the performance of MERMAID in simulation studies, and assess its robustness by conducting sensitivity analyses in a range of scenarios of model misspecifications. We apply the proposed method to analyze COVID-19 daily confirmed infection counts, PCR testing data, and serological survey data across the United States. Based on our model, we estimate an overall COVID-19 prevalence of 12.5% (ranging from 2.4% in Maine to 20.2% in New York) and overall ascertainment rate of 45.5% (ranging from 22.5% in New York to 81.3% in Rhode Island) in the US from March to December 2020.