Indiana University Indianapolis
Department of Mathematical Sciences

Statistics Seminar

12:15pm—1:15pm, Tuesday, October 08, 2024
Zoom Meeting: Meeting ID: 845 0989 4694

Speaker: Qiongshi Lu
Department of Biostatistics & Medical Informatics,
University of Wisconsin-Madison

Title: Valid inference for machine learning-assisted genetic association analysis

Abstract:

Machine learning (ML) has revolutionized analytical strategies in almost all scientific disciplines including human genetics and genomics. A rising trend in complex trait genetics research is the ML-assisted genome-wide association study (GWAS), which applies advanced ML techniques to predict phenotypes that are difficult or expensive to measure (e.g., undiagnosed diseases, imaging-derived outcomes, molecular traits in rare tissues), and then conducts GWAS on these ML-imputed outcomes. However, all existing approaches for ML-assisted GWAS treat imputed phenotypes as observed and completely neglect the inherent uncertainty associated with "black-box" ML algorithms. In this work, we first demonstrate the risk of pervasive false positive associations in existing ML-assisted GWAS. Using real data benchmarking, we find that a shocking 81% of the associations in GWAS of ML-imputed type 2 diabetes fail to replicate in the ground truth GWAS of type 2 diabetes. We then introduce POP-GWAS, a principled statistical framework that redesigns ML-assisted GWAS, ensuring valid and powerful results irrespective of the accuracy of imputation, the choice of ML algorithm, and variables used for imputation. It is a versatile tool that can account for binary phenotypes, sample relatedness, and selection bias, making it suitable for broad applications. It also only requires GWAS summary statistics as input and is computationally fast. We provide theoretical guarantee that POP-GWAS is the statistically optimal solution to ML-assisted GWAS. Using POP-GWAS, we performed the largest-to-date GWAS and rare-variant association analysis on bone mineral density (BMD) derived from dual-energy X-ray absorptiometry imaging (DXA) at 14 skeletal sites, achieving a 9.7%-50.7% gain in effective sample size compared to conventional approaches. We identified 89 novel genome-wide significant loci not previously implicated in BMD GWAS and revealed the skeletal site-specific

genetic architecture of BMD. In this talk, I will also discuss recent extensions of this framework on rare variant association analysis and beyond regression-based statistical inference.

Bio:

Dr. Qiongshi Lu received his B.S. in mathematics from Tsinghua University in 2012 and Ph.D. in biostatistics from Yale University in 2017. He is currently an Associate Professor in the Department of Biostatistics and Medical Informatics at University of Wisconsin-Madison. Dr. Lu's research focuses on developing statistical methods to study complex trait genetics. In particular, he is interested in noncoding genome annotation, genetic risk prediction, genetic correlation estimation, and gene-environment interaction.