Risk prediction: Methods, Challenges, and Opportunities

Ruowang Li
Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, California, USA
Email: ruowang.li@cshs.org

Rui Duan
Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
Email: rduan@hsph.harvard.edu

Lifang He
Department of Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania, USA
Email: lih319@lehigh.edu

Jason H. Moore
Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, California, USA
Email: jason.moore@csmc.edu

1. Introduction to the workshop

The objective of this workshop is to delve into the current and future landscape of risk prediction within the realm of disease and epidemiological research. Discussion topics encompass everything from data sources to model implementation. The workshop will feature speakers addressing commonly used data sources—genetics, imaging, clinical, and epidemiological data—in developing prediction models. Moreover, the workshop will cover model-based and post-hoc analyses, delving into biases, uncertainty quantification, model interpretation, fairness, diversity of prediction results, and the transferability and generalizability of models across different populations and datasets. The moderated discussion session will offer a future perspective on the validation and implementation of risk prediction models. The workshop will maintain a balanced focus across all stages of risk prediction model development and validation. By emphasizing a well-rounded workshop theme instead of exclusively delving into methodologies, we aim to create an environment that fosters the exchange of ideas and viewpoints among speakers and audiences.

2. Workshop Presenters

The three-hour workshop will have a total of six presentations followed by a moderated panel discussion session. The workshop speakers are:

Randi Foraker, PhD, is the Director of the Center for Population Health Informatics (CPHI) at the Institute for Informatics (I2) and a Professor of Medicine within the Division of General Medical Sciences at Washington University in St. Louis. As director of the CPHI, she aims to improve the health of the community through data and support data access, analytics, and dissemination efforts. Her own work specializes in the design of population-based studies and the integration of electronic health record data with socioeconomic indicators, and her research portfolio has been supported by a combination of governmental and industry grants and contracts. Her most recent research has focused on the application of clinical decision support to complement risk scoring in primary care, cardiology, and oncology. Dr. Foraker also serves as Director of the Public Health Data and Training Center for the Institute for Public Health. As director of the Data and Training Center, she aims to amplify public health knowledge through data sharing, strategic partnerships with the community, and the training of future public health leaders. During the COVID pandemic, she has served as PI of the COVID umbrella IRB leveraging electronic health record data at Washington University in St. Louis and works closely with investigators who conduct research using data from our COVID Data Commons, which is maintained by I2. Dr. Foraker chairs the Epidemiology Strike Force and convenes members of the St. Louis City, St. Louis County, Jefferson County, Franklin County, and St. Charles County Departments of Public Health on a weekly basis along with academic, health system, and business partners to assist with their data architecture, management, and analytic needs during the pandemic and beyond.

Yong Chen, PhD, is Professor of Biostatistics at University of Pennsylvania. He directs a Computing, Inference and Learning Lab at University of Pennsylvania (https://penncil.med.upenn.edu/about-pi/), which focuses on integrating fundamental principles and wisdoms of statistics into quantitative methods for tackling key challenges in modern biomedical data. Dr. Chen is an expert in synthesis of evidence from multiple data sources, including systematic review and meta-analysis, distributed algorithms, and data integration, with applications to comparative effectiveness studies, health policy, and precision medicine. He is also working on developing methods to deal with suboptimal data quality issues in health system data, dynamic risk prediction, pharmacovigilance, and personalized health management. He has over 100 publications in a wide spectrum of methodological and clinical areas. Dr. Chen has been principal investigator on a number of grants, including R01s from the National Library of Medicine and National Institute of Allergy and Infectious Diseases, and Improving Methods for Conducting Patient-Centered Outcomes Research grant from Patient-Centered Outcomes Research Institute. Dr. Chen received his bachelor's degree in Mathematics at the University of Science and Technology of China, Master degree in Pure Mathematics and Ph.D. in Biostatistics at the Johns Hopkins University. He is an elected fellow of the Society for Research Synthesis Methodology, and the International Statistical Institute. He is a recipient of Best Paper Award by the International Medical Informatics Association (IMIA) Yearbook Section on Clinical Research Informatics, Institute of Mathematical Statistics Travel Award, Margaret Merrell Award for excellence in research at the Johns Hopkins University, and Distinguished Faculty Award at the University of Pennsylvania.

Graciela Gonzalez-Hernandez, PhD, is Vice Chair for Research and Education in the new Department of Computational Biomedicine at Cedars-Sinai Medical Center. Prior to joining Cedars-Sinai in May 2022, Dr Gonzalez-Hernandez was an Associate Professor of Informatics in the Department of Biostatistics, Epidemiology and Informatics (DBEI) of the Perelman School of Medicine, University of Pennsylvania. She transferred her Health Language Processing (HLP) Lab to Cedars-Sinai, which focuses on natural language processing (NLP) and machine learning for knowledge discovery, extracting unstructured information from clinical records, journal articles, and social media postings to elucidate data patterns, trends, and relationships that can aid the discovery process in areas such as pharmacoepidemiology, clinical research, or public health monitoring and surveillance. Dr Gonzalez-Hernandez and her team have made available to the health research community novel approaches to complete pipelines for information extraction from different sources using NLP, such as the DeepADRMiner pipeline for extracting and normalizing adverse effects from social media – a unique end-to-end system that makes it possible to tap into the value of direct reports by patients. She has published over 220 peer-reviewed articles in prestigious journals and conferences, routinely making code and datasets available to other researchers, and ensuring reproducibility. These publications span multiple areas of Biomedical Informatics, including natural language processing, bioinformatics, biomedical ontologies, information retrieval, MS and machine learning, as well as domain-specific publications in collaboration with clinicians and epidemiologists. Her work has appeared in the top peer-reviewed journals, including Nature Digital Medicine, JAMA Network Open, Bioinformatics, BMC Bioinformatics, the Journal of the American Medical Informatics Association, and the Journal of Biomedical Informatics, among others, as well as in numerous informatics conference proceedings.

Bogdan Pasaniuc, PhD, is a professor of Computational Medicine, Human Genetics and Pathology&Laboratory Medicine at UCLA. Dr Pasaniuc develops statistical and computational methods to understand the genetic basis of disease, focusing on under-represented populations, integrative genomics, and biobank studies. Dr Pasaniuc group developed machine learning methods to integrate epigenetic profiles within trans-ancestry studies to localize disease variants and genes; his group introduced transcriptome-wide association studies (TWAS) using predicted gene expression as a principled approach to identify disease genes for many traits such as Schizophrenia, Ovarian Cancer and Prostate Cancer. Dr Pasaniuc serves as Associate Director of Population Genetics of the Institute for Precision Health at UCLA that links the genetics of more than 150k patients with their electronic health record to predict health outcomes, to stratify patients based on their genetic risk to disease and to translate genomics to the clinic. Dr Pasaniuc also serves as PI for the Center for Admixed populations and Health Equity and for the Biomedical Data Science Training Program for Precision Health Equity at UCLA.

John Witte, PhD, is serving as Vice Chair and professor in the Department of Epidemiology & Population Health, and as a professor of Biomedical Data Science and, by courtesy, of Genetics, he will also serve as a member of the Stanford Cancer Institute. Dr. Witte is an internationally recognized expert in genetic epidemiology. His scholarly contributions include deciphering the genetic and environmental basis of prostate cancer and developing widely used methods for the

genetic epidemiologic study of disease. His prostate cancer work has used comprehensive genome-wide studies of germline genetics, transcriptomics, and somatic genomics to successfully detect novel variants underlying the risk and aggressiveness of this common disease. A key aspect of this work has been distinguishing genetic factors that may drive increased prostate cancer risk and mortality among African American men. Providing an avenue to determine which men are more likely to be diagnosed with clinically relevant prostate cancer and require additional screening or specific treatment can help reduce disparities in disease prevalence and outcomes across populations. Dr. Witte has also developed novel hierarchical and polygenic risk score modeling for undertaking genetic epidemiology studies. These advances significantly improve our ability to detect disease-causing genes and to translate genetic epidemiologic findings into medical practice. Dr. Witte has received the Leadership Award from the International Genetic Epidemiology Society (highest award), and the Stephen B. Hulley Award for Excellence in Teaching. His extensive teaching portfolio includes a series of courses in genetic and molecular epidemiology. He has mentored over 50 graduate students and postdoctoral fellows, serves on the executive committees of multiple graduate programs, and has directed a National Institutes of Health funded post-doctoral training program in genetic epidemiology for over 20 years. Recently appointed to the National Cancer Institute Board of Scientific Counselors, Dr. Witte has been continuously supported by the National Institutes of Health.

Marinka Zitnik, PhD, is an Assistant Professor at Harvard University in the Department of Biomedical Informatics. Dr. Zitnik is Associate Faculty at the Kempner Institute for the Study of Natural and Artificial Intelligence, Broad Institute of MIT and Harvard, and Harvard Data Science. Dr. Zitnik investigates foundations of AI to enhance scientific discovery and facilitate individualized diagnosis and treatment in medicine. Her algorithms and methods have had a tangible impact, which has garnered interests of government, academic, and industry researchers and has put new tools in the hands of practitioners. Some of her methods are used by major biomedical institutions, including Baylor College of Medicine, Karolinska Institute, Stanford Medical School, and Massachusetts General Hospital.