Mathematics & Statistics Colloquium

Friday, December 8, 2023, 4:15pm-5:15pm

G5 Rolla Building and online via zoom[1]

Please join for refreshments and coffee (start at 4pm) before the talk!

Dr. JooChul Lee

Postdoctoral Researcher
Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania

Towards optimal sample-efficient model evaluation

Abstract. A common challenge for validating a risk prediction model using electronic health record (EHR) data is that labels for the predicted outcome are not directly available. Towards efficient and unbiased model validation, we study optimal sampling designs for efficiently labeling an informative subset of patients in an EHR cohort. Given a pre-specified number of outcome labels, our design aims to minimize the asymptotic variance of an improved inverse probability weighted estimator for predictive accuracy metrics. Implementation of the optimal sampling requires accurate risk estimates and the predictive accuracy metric of interest. We therefore propose to implement sampling in two steps. First a portion of the target number of labels is acquired by applying entropy sampling to a random subset of the cohort. These initial labels are then used to calibrate risk estimates and obtain an initial estimate of the predictive accuracy metric, which are then used to inform optimal sampling of the remaining target number of labels. The final estimate of the predictive accuracy metrics is obtained by applying the proposed estimator to the full cohort and all acquired labels pooled together. Furthermore, we address this issue by extending existing work on "Active Testing" (AT) methods which are designed to sequentially sample and label data for the evaluation pre-trained models. Application to a real EHR dataset indicate superior efficiency of the proposed sampling design and the proposed estimator.

Biographical Sketch. JooChul Lee is a postdoctoral researcher in the Department of Biostatistics, Epidemiology, and Informatics at the University of Pennsylvania. He received his Ph.D. in Statistics from the University of Connecticut in 2021 and his M.S. in Statistics from Korea University in 2013. His research focuses on sampling-efficient model development and evaluation, as well as the analysis of electronic health records data using machine learning algorithms like active learning, semi-supervised learning, and transfer learning. Additionally, his research interests include developing computational algorithms for big data, employing techniques like online updating or subsampling.

[1] Zoom meeting ID: 91804878551
password: mathstat