Advancing Responsible Statistical and AI/ML Methods for Harnessing the Power of Electronic Health Records

Qi Long

Division of Biostatistics, University of Pennsylvania

ABSTRACT

Rich electronic health records (EHR) data offer remarkable opportunities in advancing precision medicine (Orcutt et al., 2025), they also present daunting analytical challenges. Multi-modal data in EHR that are recorded at irregular time intervals with varying frequencies include structured data such as labs and vitals, codified data such as diagnosis and procedure codes, and unstructured data such as clinical notes and pathology reports. They are typically incomplete and fraught with other errors and biases. What's more, data gaps and errors in EHRs are often unequally distributed across patient groups: People with less access to care, often people with lower socioeconomic status, tend to have more incomplete data in EHRs. Such data issues, if not adequately addressed, would lead to biased results and exacerbate health inequities (Getzen et al. 2023). In this talk, I will share my research group's recent works on developing responsible statistical and AI/ML methods including large language models (LLMs) for addressing these challenges (Zhang et al., 2024; Consoli et al., 2024). Since LLMs are themselves plagued by various biases, I will also discuss our ongoing research on developing rigorous statistical and machine learning methods for mitigating pitfalls and risks of LLMs.

Keywords: AI/ML; Electronic Health Records; Large Language Models; Precision Medicine.

Back to Sessions List

Deep Survival Analysis for Competing Risk Modeling with Functional Covariates and Missing Data Imputation

Xiaofeng Wang, PhD[1], Penglei Guo, PhD,[1] Yan Zou, MS,[1] Abhijit Duggal, MD[2],
Shuaiqi Huang, PhD[1], Faming Liang, PhD[3]

[1]Department of Quantitative Health Science, Cleveland Clinic, Cleveland OH, USA,

[2]Department of Pulmonary and Critical Care Medicine, Cleveland Clinic, Cleveland OH, USA,

[3]Department of Statistics, Purdue University, Lafayette IN, USA

ABSTRACT

We introduce the Functional Competing Risk Net (FCRN), a unified deep-learning framework for discrete-time survival analysis under competing risks, seamlessly integrating functional covariates and handling missing data within an end-to-end model. By combining a micro-network Basis Layer for functional data representation with a gradient-based imputation module, FCRN simultaneously learns to impute missing values and predict event-specific hazards. Evaluated on multiple simulated datasets and a real-world ICU case study, FCRN demonstrates substantial improvements in prediction accuracy over random survival forests and traditional competing risks models. This approach advances prognostic modeling in critical care by effectively capturing dynamic risk factors and static predictors while accommodating irregular and incomplete data.

Keywords: Deep Learning; Competing Risks; Functional Data; Gradient-based Imputation

Mini-Batch Estimation for Deep Cox Models: Statistical Foundations and Practical Guidance

Ying Ding[1, *], Lang Zeng[1], Weijing Tang[2], Zhao Ren[3]

[1]*Department of Biostatistics and Health Data Science, University of Pittsburgh*

[2]*Department of Statistics and Data Science, Carnegie Mellon University*

[3]*Department of Statistics, University of Pittsburgh*

ABSTRACT

The stochastic gradient descent (SGD) algorithm has been widely used to optimize deep Cox neural network (Cox-NN) by updating model parameters using mini-batches of data. We show that SGD aims to optimize the average of mini-batch partial-likelihood, which is different from the standard partial-likelihood. This distinction requires developing new statistical properties for the global optimizer, namely, the mini-batch maximum partial-likelihood estimator (mb-MPLE). We establish that mb-MPLE for Cox-NN is consistent and achieves the optimal minimax convergence rate up to a polylogarithmic factor. For Cox regression with linear covariate effects, we further show that mb-MPLE is $\sqrt{n}$-consistent and asymptotically normal with asymptotic variance approaching the information lower bound as batch size increases, which is confirmed by simulation studies. Additionally, we offer practical guidance on using SGD, supported by theoretical analysis and numerical evidence. For Cox-NN, we demonstrate that the ratio of the learning rate to the batch size is critical in SGD dynamics, offering insight into hyperparameter tuning. For Cox regression, we characterize the iterative convergence of SGD, ensuring that the global optimizer, mb-MPLE, can be approximated with sufficiently many iterations. Finally, we demonstrate the effectiveness of mb-MPLE in a large-scale real-world application where the standard MPLE is intractable.

Keywords: Linear scaling rule; minimax rate of convergence; Stochastic gradient descent; Survival analysis

Back to Sessions List

A Deep Learning Feature Importance Test for Integrating Informative High-dimensional Biomarkers

Baiming Zou[1,2], James G. Xenakis[3], Meisheng Xiao[1], Apoena Ribeiro[4], Kimon Divaris[4], Di Wu[1,4], Fei Zou[1,5]

[1]*Department of Biostatistics, Giling School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA;*

[2]*School of Nursing, University of North Carolina, Chapel Hill, NC, USA;*

[3]*Department of Statistics, Harvard University, Cambridge, MA, USA;*

[4]*School of Dentistry, University of North Carolina, Chapel Hill, NC, USA;*

[5]*Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA.*

ABSTRACT

Many human diseases result from a complex interplay of behavioral, clinical, and molecular factors. Integrating low-dimensional behavioral and clinical features with high-dimensional molecular profiles can significantly improve disease outcome prediction and diagnosis. However, while some biomarkers are crucial, many lack informative value. To enhance prediction accuracy and understand disease mechanisms, it is essential to integrate relevant features and identify key biomarkers, separating meaningful data from noise and modeling complex associations. To address these challenges, we introduce the high-dimensional feature importance test (HdFIT) framework for machine learning models. HdFIT includes a feature screening step for dimension reduction and leverages machine learning to model complex associations between biomarkers and disease outcomes. It robustly evaluates each feature's impact. Extensive Monte Carlo experiments and a real microbiome study demonstrate HdFIT's efficacy, especially when integrated with advanced models like deep neural networks (DNN), termed HdFIT-DNN. Our framework shows significant improvements in identifying crucial features and enhancing prediction accuracy.

Keywords: Complex association, Dimension reduction, Interpretable and scalable predictive modeling, Non-parametric feature selection, Stable deep neural network

Presenter's email address: bzou@email.unc.edu

Back to Sessions List