Pitt Public Health

Biostatistics Research Day 2021
Program Book

Presented by
Department of Biostatistics
University of Pittsburgh

Thursday, March 4, 2021
11:00 AM – 5:00 PM
Graduate School of Public Health

Program At-A-Glance

Opening Session (11:00 AM – 11:15 AM)
- Opening Remarks: Jong-Hyeon Jeong, PhD (Interim Chair, BIOST)
- Welcome Speech: Maureen Lichtveld, MD, MPH (Dean, GSPH)

Student Oral Presentations (11:15 AM – 1:15 PM)
- Parallel sessions in two breakout rooms

Online Social (1:15 PM – 1:45 PM)

Faculty Presentations (1:45 PM – 2:45 PM)
- Six department faculty members will present snippets of their current research

Student Poster Presentation (2:50 PM – 3:50 PM)
- Speed plenary session followed by posters in breakout rooms

Alumnus Speaker (4:00 PM – 4:50 PM)
- Kelley M. Kidwell, PhD ('2012), Associate Professor of Biostatistics, Univ. of Michigan
- Title: *SMART Design for Treatment Effectiveness in Small Samples*

Full Program

Opening (11:00 AM – 11:15 AM) – Moderator: Abdus S. Wahed, PhD

Main session

Opening Remarks: Jong-Hyeon Jeong, PhD (Interim Chair, BIOST)

Welcome Speech: Maureen Lichtveld, MD, MPH (Dean, GSPH)

Student Oral Presentation I (11:15 AM – 1:15 PM) – Moderator: Lu Tang, PhD

Breakout room 1

11:15 – 11:30	**Xinlei Chen**, Sample Size and Power Determination in Stepped Wedge Cluster Randomized Trial Designs with Treatment Transition Periods
11:30 – 11:45	**Lingyun Lyu**, Semi-parametric Q-learning for Optimal Dynamic Treatment Regime with Right-censored Survival Outcome and Missing Treatment Data
11:45 – 12:00	**Xiaoqing Tan**, A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources
12:00 – 12:15	**Yue Wei**, Individual Treatment Effect Estimation through Machine Learning in Time-to-event Data
12:15 – 12:30	**Liwen Wu**, Interim Monitoring in Sequential Multiple Assignment Randomized Trials
12:30 – 12:45	**Qing Yin**, Mediation Analysis using Semi-parametric Shape-Restricted Regression Spline

---

Student Oral Presentation II (11:15 AM – 1:15 PM) – Moderator: Jiebiao Wang, PhD

Breakout room 2

11:15 – 11:30	**Manqi Cai**, Ensemble Estimation of Cell Type Fractions Across Various Deconvolution Approaches
11:30 – 11:45	**Yujia Li**, A Sparse Negative Binomial Mixture Model for Clustering RNA-seq Count Data

11:45 – 12:00	**Zhiyu Sui**, Discover Protein Signature of Synapse Loss During Aging Through High-throughput TMT-based Proteomics
12:00 – 12:15	**Xinjun Wang**, SECANT: a Biology-guided Semi-supervised Method for Clustering, Classification, and Annotation of Single-cell Multi-omics
12:15 – 12:30	**Molin Yue**, Cell Type Deconvolution and Cell-type Specific Differential Expression Analysis in Lung Tissue
12:30 – 12:45	**Jipeng Zhang**, GWAS for AMD Intermediate Phenotypes in AREDS
12:45 – 1:00	**Na Bo**, Genome-wide Association Analysis and Prediction Models for Hypertension on FHS Cohort

---

Online Social (1:15 PM – 1:45 PM)

Breakout rooms will be available for social interactions.

Faculty Presentation (1:45 PM – 2:45 PM) – Moderator: George C. Tseng, ScD

Main session

1:45 – 1:55	**Jeanine M. Buchanich, MEd, MPH, PhD**
1:55 – 2:05	**Ying Ding, PhD**
2:05 – 2:15	**Chaeryon Kang, PhD**
2:15 – 2:25	**Lu Tang, PhD**
2:25 – 2:35	**Jiebiao Wang, PhD**
2:35 – 2:45	**Jenna C. Carlson, PhD**

Student Poster Presentation (2:50 PM – 3:50 PM) – Moderator: Lu Tang, PhD

Speed presentation (2 min each presenter) in the main session followed by breakout rooms

A	**Tucker Harvey**, Reporting Standards and Statistical Rigor in Preclinical Animal Research
B	**Yichen Jia**, Deep Learning for Quantile Regression under Right Censoring: DeepQuantreg

C **Yang Ou**, Sensitivity Analysis of Causal Treatment Effect Estimation for Clustered Observational Data with Unmeasured Confounding

D **Xinhui Ran**, Ambient Fine Particulate Matter (PM2.5) Exposure and Incident Mild Cognitive Impairment and Dementia (2:50 PM – 3:50 PM)

E **Lang Zeng**, A Comparison and Assessment of Deep Neural Network Based Methods for Time-to-Event Data

F **Xueping Zhou**, Local-Network Guided Linear Discriminant Analysis for Cell-Type Classification Using Single-Cell RNA-Sequencing Data

---

Alumnus Speaker Presentation (4:00 PM – 4:50 PM) – Moderator: Abdus S. Wahed, PhD

Main session

Kelley M. Kidwell, PhD ('2012), Associate Professor of Biostatistics, University of Michigan
Title: *SMART Design for Treatment Effectiveness in Small Samples*

Abstract: Sequential, multiple assignment, randomized trial (SMART) designs are often motivated to identify tailored sequences of treatments or dynamic treatment regimens (DTRs) in larger samples. SMARTs employ at least two randomizations in sequence where only some groups may be re-randomized based on response or other characteristics related to previous treatment. We have turned standard SMART designs and analyses on their head, and instead of focusing on DTRs, we apply the design to small samples to obtain more information from a small sample of individuals. This talk will provide an overview of small sample SMART (snSMART) designs and primarily Bayesian methods for analyses. The differences between snSMART and SMART designs will be highlighted and methods to analyze snSMART data, calculate sample size, and add adaptive components will be presented. All of our methods are motivated by the current snSMART ARAMIS which seeks to find an effective treatment for individuals with the rare disease of isolated skin vasculitis, but the methods apply broadly to any disease affecting small numbers of individuals.

4:50 PM – 5:00 PM

Closing Remarks: Jong-Hyeon Jeong, PhD (Interim Chair, BIOST)

Biographies

Maureen Lichtveld, MD, MPH

Maureen Lichtveld, MD, MPH, is dean of the Graduate School of Public Health, where she oversees the growth and continued success of the school's seven academic departments and hundreds of students, faculty, and staff. She also serves as professor of environmental and occupational health and is the Jonas Salk Professor of Population Health. Dr. Lichtveld studies environmental public health, focusing on environmentally induced disease, health disparities, environmental health policy, disaster preparedness, public health systems, and community resilience. Her research examines the cumulative impact of chemical and non-chemical stressors on communities facing environmental health threats, disasters, and health disparities.

Kelley M. Kidwell, PhD

Kelley M. Kidwell, PhD, is an Associate Professor at the University of Michigan, School of Public Health, Department of Biostatistics. Kelley joined the University of Michigan after she earned her doctoral degree in Biostatistics from the University of Pittsburgh Graduate School of Public Health in 2012. It was in graduate school that she found her passion for clinical trial design and analysis by working on her dissertation with Abdus Wahed and through her graduate student assistantship at the National Surgical Adjuvant Breast and Bowel Project.

Kelley has continued to pursue statistical development in clinical trial design and analysis and collaborates with investigators across a wide variety of disease areas, including cancer, mental health, and rare diseases. Her research centers on the design and analysis of clinical trials, especially sequential multiple assignment randomized trials. She received a Patient Centered Outcomes Research Initiative award in 2016 to develop statistical methodology for novel trial design in rare diseases or more broadly, small samples. She was just awarded an FDA contract to continue to work in this area. Collaboratively, she has designed many trials and has been co-investigator on over 30 NIH or industry sponsored grants. She has over 100 peer-reviewed publications in both statistical and medical journals. Beyond research, Kelley enjoys teaching biostatistics to non-biostatisticians and mentoring students.

Student Abstracts

Oral Presentation Session I

Xinlei Chen, Sample Size and Power Determination in Stepped Wedge Cluster Randomized Trial Designs with Treatment Transition Periods

Determine the sample size and power in a relatively more complex SW-CRT settings where the trial has treatment transition periods by Monte Carlo simulation and compare different Generalized Linear Mixed Model (GLMM) formulations along with different estimation methods.

Lingyun Lyu, Semi-parametric Q-learning for Optimal Dynamic Treatment Regime with Right-censored Survival Outcome and Missing Treatment Data

An optimal dynamic treatment regime (DTR) is a sequence of treatment decisions that yields the best expected outcome. Limited work has been reported for estimating optimal DTRs when the outcome is survival time with right-censoring. We propose a new approach to estimate optimal DTRs in the survival setting using Q-learning, where we posit Cox Proportional Hazards (CPH) models to estimate the treatment rule for each stage, and then use the weighted hot-deck multiple imputation method to predict the optimal potential survival time for patients who did not receive optimal treatment at that stage. Furthermore, we extend the proposed method to the incomplete data setting by using inverse probability weighting and multiple imputation methods to handle missing data. Our proposed method offers several advantages. First, the hot-deck imputation method does not rely on parametric models, thus the predicted potential survival times is potentially less sensitive to model misspecification. Second, only plausible potential survival times can be imputed. Third, use of CPH models provides more flexibility regarding the shape of the baseline hazard. Extensive simulation studies were conducted to show the effectiveness of our method in identifying optimal DTR.

The methods are implemented to optimize DTRs for treatment of Leukemia from a single stage randomized study that also captured subsequent treatments.

Xiaoqing Tan, A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources

Federated learning is an appealing framework for analyzing sensitive data from distributed health data networks due to its protection of data privacy. Under this framework, data partners at local sites collaboratively build an analytical model under the orchestration of a coordinating site, while keeping the data decentralized. However, existing federated learning methods mainly assume data across sites are homogeneous samples of the global population, hence failing to properly account for the extra variability across sites in estimation and inference. Drawing on a multi-hospital electronic health records network, we develop an efficient and interpretable tree-based ensemble of personalized treatment effect estimators to join results across hospital sites, while actively modeling for the heterogeneity in data sources through site partitioning. The efficiency of our method is demonstrated by a study of causal effects of oxygen saturation on hospital mortality and backed up by comprehensive numerical results.

Yue Wei, Individual Treatment Effect Estimation through Machine Learning in Time-to-event Data

One important aspect of precision medicine is to allow physicians to choose the treatment that is most suitable for patients based on their own clinical and genetic characteristics. Different from traditional clinical studies where the focus is on estimating the average treatment effect in a representative population (usually through a well-designed clinical trial), assisting patients to shape their individualized-treatment plan requires an understanding of the heterogeneity of treatment effects from a more patient-centric view. With the increase amount of large bioinformatic datasets and the use of electronic health record data, a full picture of individuals' characteristics is forming, which also brings challenges to statistical analyses due to the complexity of the data structure. Thus, using flexible modeling techniques such as machine learning methods within the counterfactual framework shows great potential and receives much attention. In this research, we develop and apply the meta-algorithm with machine learning base learners for survival outcomes to estimate conditional average treatment effect for the observed individuals that represent the cohort with the same set of characteristics. Simulations are conducted to compare the performance of S-, T-, and X-learner paring with random survival forest or Bayesian accelerate failure time model under various conditions. We further apply the method with best performance on an Age-related macular degeneration (AMD) study and identify biomarkers that contribute to the heterogeneous treatment effects.

Liwen Wu, Interim Monitoring in Sequential Multiple Assignment Randomized Trials

A sequential multiple assignment randomized trial (SMART) facilitates comparison of multiple adaptive treatment strategies (ATSs) simultaneously. Previous studies have established a framework to test the homogeneity of multiple ATSs by a global Wald test through inverse probability weighting. SMARTs are generally lengthier than classical clinical trials due to the sequential nature of treatment randomization in multiple stages. Thus, it would be beneficial to add interim analyses allowing for early stop if overwhelming efficacy is observed. We introduce group sequential methods to SMARTs to facilitate interim monitoring based on multivariate chi-square distribution. Simulation study demonstrates that the proposed interim monitoring in SMART (IM-SMART) maintains the desired type I error and power with reduced expected sample size compared to the classical SMART. Lastly, we illustrate our method by reanalyzing a SMART assessing the effects of cognitive behavioral and physical therapies in patients with knee osteoarthritis and comorbid subsyndromal depressive symptoms.

Qing Yin, Mediation Analysis using Semi-parametric Shape-Restricted Regression Spline

Epidemiologists often build models to analyze the relationship between an exposure and a potential outcome caused by the exposure. In many cases, the exposure may not directly lead to the outcome, but instead, it induces the outcome through a process. Mediation analysis is designed to explain the causal relationship between the exposure and the outcome by examining the intermediate stage, which helps researchers understand the pathway whereby the exposure affects the outcome. The regression-based mediation analysis has been formulated and developed in the last decade, and several papers discussed the situation where the relationship between the mediator and the outcome is curvilinear. In this paper, we develop a method to analytically estimate the direct and indirect effects when we have some prior knowledge on the relationship between the mediator and the outcome

(increasing, decreasing, convex or concave) and obtain the asymptotic confidence intervals of those effects via delta method.

Oral Presentation Session II

Manqi Cai, Ensemble Estimation of Cell Type Fractions Across Various Deconvolution Approaches

Cell type fractions have been shown useful in many genomics analyses. Traditional methods for determining cell-type fractions like immunohistochemistry and flow cytometry remain costly compared to computational approaches using bulk RNA-seq data. Many computational methods, which are called cell type deconvolution, have been proposed to infer cell-type fractions from bulk transcriptomics data. However, these methods produce very different results under different settings. Motivated by Cobos et al. (2020) that performed a benchmarking study using different deconvolution procedures on simulated pseudo bulk data, we instead benchmark 12 deconvolution methods on 14 real bulk datasets using 27 reference datasets. The results show no best combination of different factors in terms of deconvolution performance. Therefore, we propose using ensemble estimation to incorporate and unify different deconvolution methods, marker gene selection procedures, data transformations, and normalizations in estimating cell-type fractions. We evaluate our method's performance on real data from different tissues and find that our method yields more stable results than existing deconvolution methods.

Yujia Li, A Sparse Negative Binomial Mixture Model for Clustering RNA-seq Count Data

Clustering with variable selection is a challenging yet critical task for modern small-n-large-p data. Existing methods based on sparse Gaussian mixture models or sparse K-means provide solutions to continuous data. With the prevalence of RNA-seq technology and lack of count data modeling for clustering, the current practice is to normalize count expression data into continuous measures and apply existing models with a Gaussian assumption. In this paper, we develop a negative binomial mixture model with lasso or fused lasso gene regularization to cluster samples (small n) with high-dimensional gene features (large p). A modied EM algorithm and Bayesian information criterion are used for inference and determining tuning parameters. The method is compared with existing methods using extensive simulations and two real transcriptomic applications in rat brain and breast cancer studies. The result shows superior performance of the proposed count data model in clustering accuracy, feature selection and biological interpretation in pathways

---

Zhiyu Sui, Discover Protein Signature of Synapse Loss During Aging Through High-throughput TMT-based Proteomics

Cognitive performance progressively declines during adults' aging. Synapse loss likely contributes to cognitive decline in normal aging. Currently, the molecular basis for declining synapse numbers in aging is unknown, precluding the development of new therapies or the intelligent application of

approved compounds. Synapse number and strength are governed by the expression, trafficking and post-translational modification of thousands of synaptic proteins. The objective of this study is to identify the synaptic protein alterations linked to aging and age-related synaptic decline that could serve as drug targets for age-related cognitive decline.

High-throughput precuneus region proteomics data was obtained through TMT (Tandem Mass Tag)-based MS (Mass Spectrometry) technology. Dendritic spine density data was collected from IHC (Immunohistochemistry) analysis. Data filtering and normalization procedures including Sample Loading normalization, Internal Reference Scaling and median normalization were applied sequentially for proteomics data. Protein-level analysis was performed to identify significantly altered proteins with aging. Then WGCNA (Weighted Gene Co-expression Network Analysis) was used to identify co-regulated protein modules. Moreover, we investigated if age-related spine density loss were associated with each protein or co-regulated module, with appropriate covariates adjusted.

Plenty synaptic protein signatures were revealed significantly correlated with aging and age-related synapse loss in human postmortem brain tissue. Recent advances in both mass spectrometry instrumentation and informatics allowed us for the investigation for high-throughput proteomics data, increasing the power to detect protein alterations as potential drug targets for age-related cognitive decline in the future.

Xinjun Wang, SECANT: a Biology-guided Semi-supervised Method for Clustering, Classification, and Annotation of Single-cell Multi-omics

The recent advance of single cell sequencing (scRNA-seq) technology such as Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) allows researchers to quantify cell surface protein abundance and RNA expression simultaneously at single cell resolution. Although CITE-seq and other similar technologies have quickly gained enormous popularity, novel methods for analyzing this new type of single cell multi-omics data are still in urgent need. A limited number of available tools utilize data-driven approach, which may undermine the biological importance of surface protein data. In this study, we developed SECANT, a biology-guided SEmi-supervised method for Clustering, classification, and ANnoTation of single-cell multi-omics. SECANT can be used to analyze CITE-seq data, or jointly analyze CITE-seq and scRNA-seq data. The novelties of SECANT include 1) using confident cell type labels identified from surface protein data as guidance for cell clustering, 2) providing general annotation of confident cell types for each cell cluster, 3) fully utilizing cells with uncertain or missing cell type labels to increase performance, and 4) accurate prediction of confident cell types identified from surface protein data for scRNA-seq data. Besides, as a model-based approach, SECANT can quantify the uncertainty of the results, and our framework can be easily extended to handle other types of multi-omics data. We successfully demonstrated the validity and advantages of SECANT via simulation studies and analysis of public and in-house real datasets. We believe this new method will greatly help researchers characterize novel cell types and make new biological discoveries using single cell multi-omics data.

Molin Yue, Cell Type Deconvolution and Cell-type Specific Differential Expression Analysis in Lung Tissue

Many computational cell type deconvolution methods have been developed recently to directly infer sample constituents from gene expression data. As a potential confounder, cell type composition could be estimated by deconvolution and adjusted in differentially expressed genes (DE) analysis to increase analysis power and reduce false positive. A key step in cell-type deconvolution is building signature matrix, which typically consists a panel of cell-type-specific genes. Nowadays, high quality signature gene expression profile in some tissues (PBMC, Brain, cancer, etc.) are available but not for lung tissue. In this study, we aim to build a high-quality signature matrix for lung/airway tissue by integrating single-cell and bulk RNA-seq data. After this matrix is built, we will deconvolute bulk RNA-seq data from our previous asthma study in lung to get cell type constituents and estimate cell-type-specific expression. The detected cell-type-specific DE genes will be compared with those DE genes identified in bulk RNA-seq data to understand the signal origin. In addition, we are able to do cell-type specific gene enrichment and pathway analysis in comparison to that from bulk RNA-seq data. Our pipeline provides a new approach for dissecting DEs from many published studies based on bulk data and prioritizing cell types and genes for functional studies.

Jipeng Zhang, GWAS for AMD Intermediate Phenotypes in AREDS

Age-related macular degeneration (AMD) is the leading cause of irreversible blindness in the U.S. Recent studies have been focusing on AMD severity scale grading, prediction of late AMD progression, and Genome-wide association studies (GWAS) of advanced AMD. In this work, we conduct the first GWAS of intermediate AMD-related phenotypes such as drusen size, drusen area, and pigmentary abnormalities, which are rarely explored previously. These intermediate phenotypes play an important role in grading severity scores and estimating future risk of advanced AMD. We performed a genome-wide association study on 4715 subjects from the Age-Related Eye Disease Study (AREDS). We identified over 10 novel genes ($P < 5 \times 10\text{-}8$) associated with these phenotypes. Among these top genes, many of them such as CFH, ARMS2 were previously identified from the GWAS of advanced AMD. We are replicating our results in UK Biobank datasets and our local cohort. Our study will provide novel insights to understand genetic basis of retinal diseases through diagnosis-relevant measurements.

Na Bo, Genome-wide Association Analysis and Prediction Models for Hypertension on FHS Cohort

High blood pressure is found to be a major risk factor of cardiovascular diseases. Many studies have explored the associations between variants and interindividual blood pressure variation through genome-wide association studies (GWAS). In this study, we conducted GWAS on multiple endpoints associated with blood pressure in the Framingham Heart Study (FHS) cohort. The FHS is a comprehensive multi-generational study, which aims to examine the epidemiology of cardiovascular disease. In our study, more than 7,000 participants (from three generations) were genotyped through Illumina 50k and Affymetrix 550k arrays. Imputation was also performed on these using the 1000G reference haplotypes. We performed GWAS on the continuous systolic and diastolic blood pressures, as well as GWAS on the binary hypertension endpoint, adjusting for sex, age, age squared, and BMI. Next, we used the top SNPs from GWAS results to build robust prediction models of hypertension, including 1) regression-based models using polygenetic scores, 2) tree-based models, and 3) neural

network prediction models. We evaluated the performance of these models and demonstrated that the top variants improved the prediction accuracy for hypotension on top of clinical risk factors.

Poster Presentation

Tucker Harvey, Reporting Standards and Statistical Rigor in Preclinical Animal Research

Preclinical studies using animal models of human diseases are usually the precursors to clinical trials for new drug treatments and other therapeutic interventions. Unfortunately, preclinical studies often suffer from poor reproducibility and a relative lack of standardization compared to RCTs. Statistical reporting standards are meant to ensure thorough and transparent reporting of findings, appropriate interpretations of results, and reproducibility of methods and statistical analyses. Examples of such guidelines do exist for preclinical research, such as ARRIVE, which intends to increase scientific quality and reporting detail in preclinical publications. However, different journals use a variety of different reporting guidelines in practice. Clinical trials tend to be held to stricter standards. One of the most common sets of such standards that includes statistical reporting items are the CONSORT guidelines, a checklist of best-practice standards for randomized controlled clinical trials. These included reporting of randomization and blinding techniques, justification for chosen sample sizes, reason for statistical methods chosen for comparison and inference, and detail regarding subgroup and adjusted analyses. For this research, preclinical publications relating to stroke and other neurodegenerative diseases in animal models were examined for whether their journal required the use of any reporting tool, and how closely they adhered to such guidelines, pertaining to statistical and study design details. It was found that not all journals require the use of reporting guidelines or use a similar set of standards. Some of the publications reviewed did not adhere to all journal-required reporting guidelines, and either omitted details pertaining to the statistical methods employed in the study or were suspected of making statistical errors. Although attempts to create standardized reporting criteria and guidelines for statistical methods and results help to increase overall quality and reproducibility, there is still work needed to enforce and hone the application of such tools in practice, so that preclinical studies can more appropriately support RCTs in standards and statistical rigor.

Yichen Jia, Deep Learning for Quantile Regression under Right Censoring: DeepQuantreg

"The computational prediction algorithm of neural network, or deep learning, has drawn much attention recently in statistics as well as in image recognition and natural language processing. Particularly in statistical application for censored survival data, the loss function used for optimization has been mainly based on the partial likelihood from Cox's model and its variations to utilize existing neural network library such as Keras, which was built upon the open source library of TensorFlow. This paper presents a novel application of the neural network to the quantile regression for survival data with right censoring, which is adjusted by the inverse of the estimated censoring distribution in the check function. The main purpose of this work is to show that the deep learning method could be flexible enough to predict nonlinear patterns more accurately compared to the traditional method, emphasizing practicality of the method for censored survival data.

Simulation studies were performed to generate nonlinear censored survival data and compare the deep learning method with the traditional quantile regression method in terms of prediction accuracy.

The proposed method is illustrated with a publicly available breast cancer data set with gene signatures."

Yang Ou, Sensitivity Analysis of Causal Treatment Effect Estimation for Clustered Observational Data with Unmeasured Confounding

Estimation of causal treatment (exposure) effect in presence of unmeasured treatment-outcome confounders often produces bias for data from an observational study. Sensitivity analysis is a useful tool for accessing how robust a treatment effect estimation with many methods developed to date. This paper proposes a new sensitivity analysis technique under clustered observational design with single study or multiple studies (meta-analysis) involved. Unlike the existing sensitivity analysis methods, our methods do not require additional assumptions on the number of unmeasured confounders, correlation between measured and unmeasured confounders, and interaction between measured confounders and the treatment. Our methods are easy to implement using the standard statistical software packages.

Xinhui Ran, Ambient Fine Particulate Matter (PM2.5) Exposure and Incident Mild Cognitive Impairment and Dementia

Poor air quality is implicated as a risk factor for cognitive impairment and dementia. However, few studies have examined these associations longitudinally in well-characterized population-based cohorts with standardized annual assessment of both mild cognitive impairment (MCI) and dementia. We aimed to determine the association between estimated ambient fine particulate matter (PM2.5) and risk of incident MCI/dementia in a post-industrial region known for historically poor air quality. The Monongahela-Youghiogheny Healthy Aging Team (MYHAT) study is an ongoing population-based cohort study of cognitive impairment in southwestern Pennsylvania, with annual assessments including a Clinical Dementia Rating (CDR). We estimated ambient PM2.5 exposure (µg/m3, single year and 5-year averages) by geocoding MYHAT participants' residential addresses to census tracts with daily PM2.5 measurements from 2002-2014 (Environmental Protection Agency air quality monitors). Using Bayesian spatial survival time-dependent models adjusted for age, sex, education, and smoking history, we examined the association between estimated PM2.5 exposure and risk of incident mild cognitive impairment (CDR=0.5) and incident dementia (CDR$\geqslant$1.0).

Lang Zeng, A Comparison and Assessment of Deep Neural Network Based Methods for Time-to-Event Data

Deep learning is a special type of machine learning method for learning complex data through neural networks. Recent methodological development of deep neural network (DNN) and technological advances in generating large-scale data enable researchers unprecedented opportunities to use DNN to analyze the high-dimensional data with time-to-event outcomes. In this work, we compared six recently developed DNN survival methods (DNNsurv, Nnet-Survival, DeepHit, DeepSurv, Cox-nnet, and CapSurv) regarding their assumptions, data input/output, neural network design, loss function, and applications. Among those methods, we applied three of them (DNNsurv, DeepSurv and Nnet-Survival) to the Age-Related Eye Disease Study (AREDS) data to predict the risk of Age-related Macular Degeneration (AMD) progression with clinical and genetic features. Prognostic accuracies of methods were compared using two different metrics: C-index and Brier score. In the future, motivated

by the method with the best performance in this preliminary analysis, we will further develop a robust multi-layer neural network survival model to predict AMD progression risk profile over time using both GWAS and longitudinal fundus images data.

Xueping Zhou, Local-Network Guided Linear Discriminant Analysis for Cell-Type Classification Using Single-Cell RNA-Sequencing Data

Single-cell RNA sequencing (scRNA-seq) technology has yielded massive transcriptional profiles of individual cells. Such rich data provide the opportunities to detect the important transcriptional signals and predict cell types. With the high-dimensional scRNA-seq data, the most predictive genes need to be first selected. Most of the current machine learning techniques only select genes that are strongly correlated with the outcome cell type. However, many genes even though have marginally weak correlations with the outcome cell types, may execute a strong predictive effect on the cell type classification. Here, we employ the local-network guided multiclass linear discriminant analysis (mLDA), which incorporates the weak signals into predictions, to predict cell types using scRNA-seq data from human peripheral blood mononuclear cells. The results show that the prediction accuracy is significantly improved when the predictive weak signals are included.