Zhengwei Song, M.S. Personal Website
Eli Lilly and Company
Lilly Corporate Center, Indianapolis, IN 46225
Phone: (332) 256-6895
Email: zs2539@caa.columbia.edu

EDUCATION

Columbia University, New York, NY 09/2022 – 05/2024
M.S. in Biostatistics, GPA: 4.03/4

- Relevant Coursework: Advanced Probability, Advanced Statistical Learning & Data Mining, Advanced Statistical Computing, Biostatistical Methods, Data Sciences, Survival Analysis, Randomized Clinical Trial, Epidemiology

University of Manchester, Manchester, United Kingdom 09/2019 – 06/2021
B.Sc. in Mathematics and Statistics (1st honor), GPA: 72.07/100

- Relevant Coursework: Real Analysis, Statistical Inference, Markov Chain, Martingales

Shandong University, Jinan, China 09/2017 – 06/2021
B.S. in Mathematics, GPA: 4.02/5 (top 20%-35%)

- Relevant Coursework: Adv. Algebra, Mathematical Analyses

PUBLICATIONS

1. Lee AJ, Cui Z*, **Song Z***, Reyes-Dumeyer D, De Jager PL, Bennett DA, Schneider JA, Menon V, Wang Y, Lantigua RA, Medrano M, Rivera D, Jiménez-Velázquez IZ, Kukull WA, Brickman AM, Manly JJ, Tosto G, Kizil C, Farrer LA, Mez J, Chung J, Vardarajan BN, Mayeux R. Genome-wide gene-based study in multi-ethnic cohorts identifies genes that interact with vascular risk factors in Alzheimer's Disease. In preparation for ***JAMA Neurology***. *Contributed equally.

2. **Song Z**, Lee AJ. Integrated Transcriptomics and Epigenomics Analysis Identifies Molecular Subtypes of Alzheimer's Disease. In preparation for ***Neurology***.

RESEARCH EXPERIENCE

Adaptive Treatment Design and Multi-Armed Bandit (MAB) Optimization 08/2024 – Present
Advisor: Dr. Min Qian, Dr. Bin Cheng, Department of Biostatistics, Columbia University

- Examine adaptive allocation rules in multi-armed bandit problems, focusing on maximizing cumulative rewards through upper confidence bound (UCB) algorithms.
- Develop a novel framework combining MAB algorithms with online false discovery rate (FDR) control to enhance efficiency and accuracy in sequential A/B testing scenarios.
- Re-define the null hypotheses in MAB instances, derive less-conservative always-valid sequential p-values for continuous monitoring, and integrate FDR rejection thresholds to optimize sample complexity and statistical power.

Data-driven dynamic modeling for Alzheimer's disease (AD) progression 05/2024 – Present
Advisor: Dr. Ying Wei, Dr. Tianying Wang, Department of Biostatistics, Columbia University

- Design and implement a multivariate dynamic model using ordinary differential equations to jointly model rates of change in biomarkers and cognitive tests, to analyze the temporal evolution of biomarkers and cognitive outcomes, addressing the heterogeneity in sporadic AD populations.

- Apply the model to longitudinal data from the Researcher's Data Dictionary-Genetic (RDD-Gen), illustrating biomarker progression patterns and predicting time to conversion from mild cognitive impairment to dementia with interpretable and flexible outputs.

Enhancing Breast Cancer Risk Prediction in Hispanic Women Through Transfer Learning 01/2024 – 05/2024
Advisor: Dr. Tian Gu, Department of Biostatistics, Columbia University

- Applied single nucleotide polymorphism (SNP) clumping to develop and test improved breast cancer risk prediction models for the Hispanic population.
- Utilized Elastic Net on SNP selection and determined 347 relevant genetic markers for risk prediction.
- Used innovative angle-based transfer learning with genetic data from three ancestral populations and improved predictive accuracy of up to 100%, comparing them to established benchmark methods.

Broadening gene discovery for AD by incorporating additional cardiovascular and cerebrovascular risk factors (CVRFs) and examining the multi-omics profiles of genes to unravel their mechanisms and causal pathways 03/2023 – 05/2024
Advisor: Dr. Annie Lee, Department of Neurology, Columbia University

1. *Multi-omics integration via Similarity Network Fusion (SNF) for identifying molecular subtypes of aging, for effective treatments with cerebrovascular factors*
   - Utilized Random Forest for missing data imputation to ensure the similarity networks were based on comprehensive and representative data
   - Integrated RNA-seq (1092 samples, 18629 genes), proteomics (400 samples, 8817 proteins), and DNA Methylation (704 samples, 420132 CpG sites) data by applying Similarity Network Fusion, identifying molecular subtypes of Alzheimer's patients via spectral clustering to comprehensively understand AD's heterogeneity across patient subtypes
   - Tested associations betweem cerebrovascular factors (e.g. infarctions in 45 brain regions) and subtypes, identified five significant infarct regions that could inform personalized treatments

2. *Causal Mediation Analysis and Trait Analysis on CVRFs interacted, AD risk altered genes, to quantify the involvement of gene expression in brain pathology in AD*
   - Adjusted technical variables (e.g. batch) identified by a forward selection approach using the voom/limma pipeline in RNA-seq data to ensure analysis results were not biased due to variations across different technical variables
   - Used adaptive gene-environment interaction (aGE) test to test for the genes that interacted with CVRFs to alter AD risk
   - Analyzed the relationships among 45 cerebral infarctions, Alzheimer's disease, Amyloid-β, and tau to see if gene expression was influenced by amyloid or phosphorylated tau.

3. *Multi-ancestry mQTL analysis of human brain identifies candidate causal variants for brain-related traits*
   - Divided participants as vascular, neurodegenerative and mixed group according to pathological data to identify participants without neurodegenerative pathologies, according to clinician's evaluation
   - Conducted a large-scale multi-ancestry methylation quantitative trait loci (mQTL) analysis using ROSMAP data, integrating 2,385 DNA methylation sequencing samples from 2,119 donors, including 474 non-European individuals.
   - Integrated mQTL and GWAS data to perform joint statistical fine-mapping, pinpointing over 100 unique candidate pathogenic variants associated with brain-related traits.
   - Identified dozens of unique genes driving these variants, uncovering regulatory mechanisms underlying vascular risk, neurodegenerative decline, and Alzheimer's disease.

Black-Scholes Pricing Model Data Simulation by Multilevel Monte-Carlo Method 09/2020 – 05/2021
Advisor: Dr. Jianliang Chen, School of Mathematics, Shandong University

- Executed Monte-Carlo simulations for path-dependent option pricing based on Weiner process models
- Developed R scripts for Asian Option pricing, performed asset path averaging, payoff calculation, and variance reduction in multilevel Monte Carlo methods

Edible Tableware based on Finite Element Analysis 03/2019 – 06/2019
Advisor: Dr. Song Yu, Department of Engineering Mechanics, Shandong University

- Led a team of five with diverse academic backgrounds and secured full funding (around $900)
- Designed and produced a chopstick-like mold by SolidWorks software according to finite element analysis theories
- Connected and partnered with local restaurants and bars for testing mechanical characteristics

PROFESSIONAL EXPERIENCE

Senior Computational Statistician, Eli Lilly and Company, Indianapolis, IN 07/2024 – Present

- Review protocols and statistical analysis plans related to the development of new drugs for immunological diseases, conduct data analysis and TFLs, write analysis reports, and address regulatory issues.
- Collaborate with the data management team to implement data quality assurance strategies, follow the CDISC clinical trial standard, derive SDTM and ADaM data formats, and write data specifications.
- Study relevant literature in the therapeutic area and regularly present advanced clinical trial analysis tools, such as robust data imputation methods and adaptive randomization.

Biostatistics Intern, Roche Holding AG, Shanghai, China 04/2022 – 09/2022

- Developed statistical methods for analyzing clinical trial data, including the development of novel approaches to address specific research questions and issues with existing methodologies
- Collaborated with 10 medical team members and provided statistical support in phase 4 clinical trials by co-developing analytical plans, performing analyses, interpreting results, and summarizing findings into concise reports that are understandable to non-statisticians
- Co-developed an R package (impost) of linear mixed effects models for the tumor size over time by Bayesian inference using Hamiltonian Monte Carlo method (Stan)

Data Analyst Intern, Sina Corporation, Beijing, China 10/2021 – 04/2022

- Scraped & wrangled Weibo user data, and created visualization (user portraits) for rankings in the entertainment operations
- Presented final statistics for several popular TV series, variety shows, and documentaries, to provide data support for social media influencers and internal operations
- Maintained data warehouse services under the Hive SQL environment

EDUCATIONAL CONTRIBUTIONS

TEACHING

- Teaching Assistant, "**Data Science II**" (P8106), Department of Biostatistics, Columbia University Spring 2024
  Instructor: Dr. Yifei Sun

- Teaching Assistant, "**Biostatistical Methods I**" (P8130), Department of Biostatistics, Columbia University Fall 2023
  Instructor: Dr. Molei Liu

- Teaching Assistant, "**Introduction to Mathematical Statistics**" (P8107), Fall 2023
  Department of Biostatistics, Columbia University
  Instructor: Dr. R. Todd Ogden

MENTORSHIP

- Mentoring, Danielle Savellano, Sophomore in Biology, Allegheny College Summer 2023
- Mentoring, Kayla Scott-McDowell, Junior in Biochemistry & Sociology, Summer 2023
  Mt. Holyoke College
  Mentoring two undergraduate students from Biostatistics Epidemiology Summer Training (BEST) Diversity Program belonging to **Summer Health Professions Education Program (SHPEP)**, Columbia University
  Advisor: Dr. Annie Lee

PRESENTATIONS

1. Lee AJ, Cui Z, **Song Z**, Reyes-Dumeyer D, De Jager PL, Bennett DA, Schneider JA, Menon V, Wang Y, Lantigua RA, Medrano M, Rivera Mejia D, Jiménez-Velázquez IZ, Kukull WA, Biber SA, Brickman AM, Tosto G, Kizil C, Farrer LA, Mez J, Chung J, Vardarajan BN, Mayeux R. "Multi-ancestry Genome-wide Gene-Vascular Risk Factors Interaction Analyses in Alzheimer's Disease." Alzheimer's Association International Conference (AAIC), Philadelphia, PA, *contributed.* (08/2024)

2. **Song Z**, Gu T. "Enhancing Breast Cancer Risk Prediction in Hispanic Women Through Transfer Learning." Biostatistics Practicum/APEX Symposium, Department of Biostatistics, Columbia University, NY (04/2024)

3. **Song Z**, Cui Z, Lee AJ. "Broadening Gene Discovery for Alzheimer's Disease." Columbia University Data Science Day, New York, NY (04/2024)

4. Lee AJ, Cui Z, **Song Z**, Reyes-Dumeyer D, De Jager PL, Bennett DA, Schneider JA, Menon V, Wang Y, Lantigua RA, Medrano M, Rivera D, Jiménez-Velázquez IZ, Kukull WA, Brickman AM, Manly JJ, Tosto G, Kizil C, Farrer LA, Mez J, Chung J, Vardarajan BN, Mayeux R. "Genome-wide Gene-based study in Multi-ethnic Cohorts identifies Genes that Interact with Vascular Risk Factors in Alzheimer's Disease." Alzheimer's & Parkinson's Diseases and related neurological disorders (AD/PD), Lisbon, Portugal, *contributed.* (03/2024)

HONORS & AWARDS

1. Sunshine Award (Top 44/3692), "AAAI-2022 AIC Phase VIII: Data-Centric Robust Learning on ML Models." Jointly organized by Alibaba Security, Tsinghua University, RealAI, and Association for the Advancement of Artificial Intelligence (AAAI) (01/2022)

2. Annual Outstanding Individual (Top 1/25) from project "Edible Tableware based on Finite Element Analysis." Enactus (12/2018)

3. Third-level Academic Scholarship (Top 30% in GPA), Shandong University (10/2018)

SKILLS

Environments: R (tidyverse, gtsummary, caret, survival, lme4, gee, httr, bioconductor, etc.) | RShiny | SAS | SQL | PASS | C | MATLAB | Unix | Microsoft Office | AutoCAD
Test: t, z, ANOVA, chi-squared, Fisher's exact, McNemar's, Log-rank, sign, Wilcoxon signed-rank & rank-sum, etc.

Modeling: linear, generalized linear (logistic, Poisson, etc.), weighted least squares, mixed effect, GEE, survival (Cox, Stratified PH, AFT),

Machine Learning: decision tree, random forest, boosting, K-NN, cubic splines, local regression, GAM, MARS, LDA, QDA, NB, SVM, clustering (K-means, Hierarchical, spectral, etc.), PCA, LASSO, Elastic net, Ridge, PCR, PLS, cross validations

Monte Carlo simulation, optimization (Newton-Raphson), EM, bootstrapping, Markov Chain Monte Carlo