Bingkun Luo

+1(217) 4179664 | bl2789@caa.columbia.edu | luo.bingkun@outlook.com

Objective

I am actively seeking a research opportunity to apply diverse methods such as missing data imputation, clinical trial design, statistical association, and machine learning/deep learning. Additionally, I am keen to develop skills in manuscript writing and editing.

Education

Columbia University, Mailman School of Public Health — New York, NY
Master of Science in Biostatistics
Graduated: May 2021
Relevant Courses: Data Mining, Relational Database & SQL Programming, Data Science, Longitudinal Data, Statistical Inference, Biostatistical Methods I & II, The Latent-Variable Structure & Modeling

University of Illinois at Urbana-Champaign — Champaign-Urbana, IL
Bachelor of Science in Actuarial Science and Statistics (Highest Distinction)
Graduated: May 2019
Relevant Courses: Applied Regression and Design, Applied Bayesian Analysis, Machine Learning, Probability, Survival Analysis, Linear Algebra

Skills

- **R, Python, Git, SQL**, MySQL, Cloud Bigdata Query, Genomics Data Query, SAS, AWS, VBA, PowerShell

Experience

Amgen — Clinical Biomarkers and Diagnostics (Remote, United States)
Data Scientist (Oct 2021 – June 2024)

- Biomarker ETL (Python): Led the ETL process for diverse biomarker assay data (e.g., blood samples, circulating tumor cells). Standardized clinical data integration across 13 programs, improving biomarker assay integration and harmonization.
- Association Pipeline & Visualization (R, Bash): Designed and maintained an automated association pipeline for various indication, delivering dynamic statistical insights for analytes (proteomics, RNA, and DNA). Produced over 400 reports for Blincyto, Tarlatamab, BiTE with standardized quality control procedure, providing systematic endpoint comparisons across baseline and various on-treatment settings.
- Machine Learning for Predictive Modeling (Python, R): Developed machine learning models (Random Forest, XGBoost) to predict adverse events. Conducted data preprocessing, feature selection (Lasso Regression, Elastic Net), and constructed performance assessment reports integrating key biomarker features.
- Survival Analysis & Visualization (R, Bash): Implemented survival analysis for clinical biomarker data, generating Kaplan-Meier curves, log-rank test statistics, and forest plots with hazard ratios to support clinical endpoints. Produced summary reports for stakeholder decision-making.
- Biomarker & Diagnostic Development (R, Bash): Collaborated in optimizing the biomarker analysis for Tezepelumab Phase 2b trials. Stratified patients using inflammatory cytokines and anti-IgE therapy, applying Linear Mixed Models to assess on-treatment effects.

Columbia University Irving Medical Center — Obstetrics & Gynecology (New York, United States)
Research Assistant (July 2020)

- Processed preterm birth variant datasets using bedtools and R. Conducted pathway analysis and statistical testing for tissue-specific SNPs in reproductive tissues.
- Utilized HiC and ATAC-seq data to analyze promoter-promoter interactions in target gene region. Applied Fisher's exact test for enrichment analysis in eQTL data.

Public Health Capabilities

- Survival Analysis: Expertise in applying censored data methodologies such as Kaplan-Meier curves, log-rank tests, and forest plots to evaluate treatment outcomes in clinical trials.
- Genetic and Molecular Epidemiology: Proficient in handling and analyzing large-scale genetic datasets (e.g., UK Biobank) to investigate genetic variants associated with diseases, including preterm birth and aging-related conditions.
- Clinical Trials: Experienced in clinical trial design and statistical analysis, including patient stratification, safety/efficacy evaluation, biomarker sample management, optimizing biomarker analysis plans for global trials.
- Applied Analytic Methods: Extensive application of machine learning (Random Forest, XGBoost) and statistical modeling techniques (Lasso, Elastic Net) for predictive modeling and feature selection in clinical and epidemiological data.
- Big Data Analytics: Skilled in handling large datasets (10,000+ observations), conducting data imputation, and utilizing tools like R Shiny for interactive visualization and presentation of public health data.

Conference Presentation

Bingkun Luo, Bharat Panwar, Clinical Biomarker and Diagnostics & Oncology & General Medicine (May 2023). Integrated Al/ML Framework for Biomarker-driven Analysis in Clinical Trials for NSCLC Treatment. Poster presented at Amgen data science symposium.

- *Hypothesis*: AI/ML analytics platform can identify composite signatures of treatment response from high-content biomarker data and clinical metadata in non-small cell lung cancer (NSCLC) patients, aiding in personalized therapy strategies.
- *Methods:* The platform utilized non-parametric statistical tests, machine learning methods with feature elimination, and Shapley values to identify significant genes, detect outlier patients, and provide explanations for their biological significance.
- *Result*: Discovered a 4-mutational gene signature associated with higher overall response rate (ORR) in Sotorasib-treated NSCLC patients using baseline training and validation RNA-seq data from the CodeBreaK dataset, demonstrating its potential for guiding treatment decisions.

Selected Projects

Big Data Analysis of Accidental Drug Uses (Columbia University, 2019)

- Processed data on drug-related deaths (10k+ observations). Developed machine learning models to predict trends in drug abuse across the US. Built an R Shiny app to facilitate exploratory data analysis and interactive visualizations.

Hierarchical Bayesian Modeling for Suicide Rates (Columbia University, 2019)

- Cleaned and analyzed suicide prevention datasets (27,820 observations). Applied Markov Chain Monte Carlo sampling to establish a hierarchical log-linear model predicting suicide rates per 100,000 people across countries.