Zhongyi (James) Guo

1300 Oak Creek Dr., Palo Alto, CA (open to remote or relocate) | guozy@stanford.edu
LinkedIn | GitHub | Personal Website

EDUCATION

Stanford University **Stanford, CA**

M.S. Epidemiology & Clinical Research (*GPA: 3.91/4.00*) *(expected) 06/2025*

Relevant Coursework: Epidemiological Methods & Study Design, Machine Learning for Biomedical Data Fusion, Causal Inference, Data Analysis using SAS, Biostatistics: Analysis of Discrete Data, Genetic Epidemiology

Cornell University **Ithaca, NY**

B.S. Biometry & Statistics, Biological Sciences (GPA: 3.57/4.30) *05/2023*

Relevant Coursework: Data Mining & Machine Learning, Data Science, Linear Algebra, Statistical Computing, Object-Oriented Programming using Java and Python, Theory of Statistics, Probability Models & Inference

Academic Honors: ***Cum Laude, Dean's List***

SKILLS

- Programming: R, Python, SAS (Base certified), Java, Swift, SQL, UNIX/Linux, LaTeX, HTML, CSS, JavaScript.
- Core Skills: Data Science, Statistical Modeling, Multi-omics Analysis, Machine Learning, Communication, Detail-Orientation, Innovation, Curiosity, Time Management, Literature Review, Excel, Word, PowerPoint.

RELEVANT RESEARCH EXPERIENCE

Graduate Research Assistant at Graff Lab **Remote**

P.I.: Dr. Rebecca Graff, Department of Epidemiology and Biostatistics, UCSF *11/2023 – Present*

- Utilized ChemRICH to study ethnic disparities between black and white men in a prostate cancer pilot study, interpreting chemical similarity enrichment analysis results using R.
- Developed a web-scraping tool to retrieve Compound IDs and SMILES from PubChem using R.
- Replicated and validated t-test results (mean values and p-values) along with Storey's q-value using the False Discovery Rate (FDR) approach based on Metabolon's report on metabolomics data in R.

RELEVANT WORK EXPERIENCE

Tencent **Remote**

Data Analyst Project Intern *07/2021 – 09/2021*

- Extracted e-commerce sales statistics using Python web scraping and using SQL on internal databases.
- Developed predictive machine learning models for forecasting sales trends using Python Sklearn, analyzing customer shopping patterns across product categories, and built multiple linear regression models.
- Refined marketing team's strategies by presenting my elegant and informative visualization, along with model building, in my data analysis report created using Python in Jupyter Notebook.

RELEVANT PROJECT EXPERIENCE

R package: hurdatPro *05/2023*

- Developed an R package in tar.gz to analyze Atlantic tropical cyclone activities through collaboration.
- Cleaned data, designed functions for storm plotting (track, position, size), identified U.S. landfalls, computed storm accumulated cyclone energy, and implemented unit tests using testthat.

Causal Effect of Type II Diabetes on Prostate Cancer in the East Asian Population *05/2022 – 12/2022*

- Performed two-sample Mendelian randomization with the inverse variance weighted method while using MR Egger and weighted median methods as sensitivity analysis on genetic-level data.
- Identified proxy SNPs in linkage disequilibrium (r2 > 0.8) and obtained OR = 0.76, 95% CI = [0.76, 0.89], P-value = 2.26 * 10-6 and similar results in sensitivity analysis.
- Concluded that Type II diabetes has a negative causal effect on prostate cancer using genetic evidence.

GWAS Study: Analysis of Citrulline Levels and Chronic Kidney Disease *05/2022*

- Performed GWAS analysis on citrulline levels and chronic kidney disease data using two PCs obtained from PCA as covariates on genotype data, and Bonferroni correction to reduce Type I error.
- Identified 2 significant SNPs from Manhattan plot with 2 covariates included and interpreted the influence of linkage disequilibrium on the result.

Weather Data Analysis in Ithaca, NY (from 01/2021 to 04/2022) *03/2022 – 04/2022*

- Built a Logistic Regression model and a K-Nearest Neighbors (KNN) model to forecast snow in Ithaca, NY, based on daily temperature range using train-test split after data cleaning & exploratory data analysis.
- Reached the model accuracy at 0.809 for the Logistic Regression and 0.786 for the KNN with k = 10 and plotted confusion matrices for two models' tuning & validating and error analysis.

α-helix or not? *12/2021*

- Performed feature engineering on the training set by averaging each feature of 5 neighboring amino acids and removed redundant features measured by correlation coefficient.
- Trained binary classifiers (Logistic Regression, Decision Tree Regressor, and Random Forest models) using Sklearn in Python to predict α-helix or not using features derived from primary structures of proteins.
- Tuned the maximum number of iterations using random search method to optimize the Logistic Regression model, and conducted cross-validation and examined the model accuracy (AUROC = 0.625).

Salaries in Big Techs *09/2021 – 12/2021*

- Built a multiple linear regression model to predict total yearly salaries based on employee features, including years of experience, gender, race, education, etc., for tech companies in US and overseas.
- Established 3 equations for users to optimize their incomes by predicting total yearly salary in the U.S.

TEACHING EXPERIENCE

Cornell University **Ithaca, NY**

Beta Tester & Teaching Assistant, Introduction to Data Science *01/2023 – 05/2023*
Grader, Probability Models & Inference *08/2022 – 12/2022*
Teaching Assistant, Laboratory in Genetics and Genomics *01/2021 – 05/2021*

RELEVANT EXTRACURRICULAR EXPERIENCE

Community HealthEd **Remote**

Education Branch – Scientific Review Editor *03/2022 – 05/2023*

- Performed literature review on each paper/website cited in the articles to validate accuracy of citations.
- Revised articles to remove technical jargon while retaining clear concise plain language to the public audience as newsletters, primarily focusing on maternal, prenatal, neurological & psychiatric health.