Junzhe Shao

Berkeley, CA | (+1) 858-766-1048 | junzhe_shao@berkeley.edu | https://junzheshao98.github.io/homepage/

Education

University of California, Berkeley Sep 2023 – 2026 (Expected)
Ph.D. Candidate in Biostatistics

Columbia University Sep 2021 – May 2023
M.S. in Biostatistics

Peking University Sep 2016 – Jun 2021
B.S. in Biological Science; Minor in Physics

University of California, San Diego Sep 2019 – Sep 2020
Visiting Student

Work Experience

Internship, Product Development Data Science Jun 2025 – Present
Genentech, South San Francisco, CA

- Developed a bias-corrected Augmented Inverse Probability Weighting (AIPW) estimator for longitudinal external-control studies, leveraging double machine learning and semiparametric efficiency theory to improve causal effect estimation.
- Built reproducible R simulation pipelines to benchmark performance and drafted the methods manuscript.
- Applied and validated on real clinical-trial data, delivering study-ready analyses.

Research Experience

Bias Corrected AIPW Estimator for External Control in Longitudinal RCTs May 2025-Present

- Developed a bias-corrected augmented inverse probability weighting (AIPW) estimator that leverages external control arms with longitudinal data.
- Established bias correction at a rate governed by the number of time points, ensuring valid type I error control compared to naïve direct pooling approaches. Provided non-asymptotic guarantees.
- Demonstrated higher statistical power compared to RCT-only estimators and other covariates adjustment methods.

Nonstationary Adaptive A/B Testing with State-space Models May 2025-Jan 2026

- Proposed a state-space modeling framework for adaptive AB testing in nonstationary environments with evolving treatment effects.
- Developed a Neyman-style optimal allocation policy via Kalman filtering and established asymptotic variance bounds for ATE estimator and provide anytime valid confidence interval.

Regularized Frameworks for Adaptive Enrichment Designs Jun 2023 – Dec 2025

- Proposed a unified adaptive enrichment framework to maximize power for testing average treatment effects.
- Cast the problem as constrained optimization to unify different objectives.
- Calibrated selection bias and provided valid inference throughout the process.

Enhanced Large Language Model for Predicting HIV Care Disengagement Dec 2024 – April 2025

- Fine-tuned LLaMA 3.1 with over 4.8 million EMR records from Tanzania to predict disengagement from HIV care, ART non-adherence, and adverse outcomes.
- Demonstrated superior predictive performance compared to supervised ML models and zero-shot LLMs, with strong internal and external validation results.
- Provided clinically interpretable insights; expert evaluation confirmed the model's reasoning was clinically relevant in over 90% of aligned cases.

Causal Inference for Non-Stationary Time Series Data Nov 2021 – Present

- Proposed a parametric generalization of synthetic control with time-varying, non-permutation-invariant weights.
- Applied the approach to study the impact of a mandatory certificate on COVID-19 vaccine compliance.
- Developing the R package `SSMimpute` for state-space-model multiple imputation in non-stationary multivariate time series.

Adaptive Design for Randomized Trials under High-dimensional Setting Jun 2022 – Present

- Investigated theoretical properties of covariate-balancing methods for randomized clinical trials under many covariates.
- Proved that a balancing criterion can be achieved asymptotically when the number of covariates is a vanishing proportion of the number of patients.

Publications & Manuscripts in Preparation

Junzhe Shao, Waverley Wei, Jingshen Wang. "*Adaptive A/B Testing under Nonstationary Dynamics using State-Space Models*", *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

Junzhe Shao, Aibo Gong, Juan Shen, Waverley Wei. "*Regularized Frameworks for Flexible and Valid Adaptive Enrichment Designs*", Under review at *Statistica Sinica*, 2025.

Waverly Wei[†], **Junzhe Shao**[†], Rita Qiuran Lyu, Rebecca Hemono, Xinwei Ma et al. "*Enhanced language models for predicting and understanding HIV care disengagement: a case study in Tanzania*", *npj Digital Medicine*, 2026.

Junzhe Shao, Mingzhang Yin, Xiaoxuan Cai, Linda Valeri. "*Generalized Synthetic Control Method with State-space Model*", short version accepted at NeurIPS 2022 Workshop for Causality for realworld impact. Oral Presentation at American Causal Inference Conference (ACIC) 2023, Manuscript in preparation.

Alton Barbehenn, Lei Shi, **Junzhe Shao**, et al. "*Rapid Biphasic Decay of Intact and Defective HIV DNA Reservoir During Acute Treated HIV Disease*", *Nature Communications*, 2024.

Mona Alotaibi[†], **Junzhe Shao**[†], Michael W. Pauciulo, et al. "*Metabolomic Profiles Differentiate Scleroderma-PAH from Idiopathic PAH and Correspond with worsened Functional Capacity*", *Chest*, 2022.

[†] Co-first authors.

Skills Summary

Programming: R, Python, SQL
Software/Tools: LaTeX, Git, Bash, Slurm
Languages: Mandarin (Native), English (Proficient)