SHIYU WAN

+1(984) 215-8034 ◇ Seattle, WA
shiyu31@uw.edu ◇ Linkedin ◇ Google Scholar

SUMMARY

PhD candidate in Biostatistics at UW (qualifying exam passed Aug 2025) with an M.S. in Biostatistics from UNC and dual B.Med/B.Econ from Peking University. I develop estimands and inference for master protocol trials and build deep learning approaches for genomic prediction and survival analysis, including supervised fine-tuning of genomic foundation models. My work aims to deliver rigorous, interpretable methodology that advances clinical trials and population-scale genomics.

EDUCATION

Doctor of Philosophy in Biostatistics, University of Washington, Seattle — Expected 06/2028
Relevant Coursework: Advanced Theory of Statistical Inference I & II & III, Advanced Regression Methods I & II & III, Design of Medical Studies.
Passed PhD Qualifying Examination in August 2025

Master of Science in Biostatistics, University of North Carolina at Chapel Hill — 08/2022 - 05/2024
Relevant Coursework: Advanced Probability and Statistical Inference I & II, Linear Model, Statistical Learning and Precision Medicine, Design and Analysis of Clinical Trials, Statistical Computing, Optimization.

Bachelor of Medicine in Preventive Medicine, Peking University — 09/2017 - 06/2022
Relevant Coursework: Medical Genetics, Surgery, Internal Medicine, Clinical Practice, Epidemiology.

Bachelor of Economics in Economics, Peking University — 09/2018 - 06/2022
Relevant Coursework: Advanced Econometrics I & II, Time Series Econometrics.

EXPERTISE AND SKILLS

- **Research Areas:** Deep learning and statistical methods for genomics and survival analysis; causal inference in master-protocol trials.
- **Methodologies:** Genomic foundation models (Enformer, Borzoi, Nucleotide Transformer); survival deep neural networks; causal inference and semiparametric efficiency theory.
- **Technical Skills:** Python (PyTorch, PyTorch-Lightning, scikit-learn), R (Bioconductor, survival analysis, causal inference packages, tidyverse), High-Performance Computing (Slurm, GPU clusters)

SELECTED PUBLICATIONS

1. **Wan S**, Rojas-Rueda D, Pretty J, Roscoe C, James P, Ji JS. Greenspace and mortality in the U.K. Biobank: Longitudinal cohort analysis of socio-economic, environmental, and biomarker pathways. *SSM Popul Health*, 2022;19:101194. doi:10.1016/j.ssmph.2022.101194
2. **Wan S**, Tao L, Liu M, Liu J. Prevalence of toothache in Chinese adults aged 65 years and above. *Community Dent Oral Epidemiol*, 2021;49(6):522-532. doi:10.1111/cdoe.12640
3. Zou B, Mi X, **Wan S**, Xenakis J, Wu D, Hu J, Zou F. A Deep Neural Network Two-part Model and Feature Importance Test for Semi-continuous Data. *Annals of Applied Statistics*. 2025; 19(2): 1314-1331.

WORKING PAPERS

1. **Wan S**, Mi X, Zou F, Zou B. An Interpretable Deep Learning Framework for Biomarker Discovery in Complex Disease Survival Outcomes, doi: 10.1101/2025.09.30.679415
2. **Shiyu Wan**, Yuhan Qian, Nicole Mayer-Hamblet, Patrick J. Heagerty and Ting Ye. Re-Enrollment in Master Protocol trials. Manuscript in preparation.

RESEARCH EXPERIENCE

Robust Estimation of Treatment Effect in Master Protocol Trials — 01/2025 - Present
Independent Research; Advisor: Professor Ting Ye, UW Seattle — *Seattle, WA*

- Defined estimands for master protocol trials with participant re-enrollment; Analyzed the statistical properties of classical causal inference methods; derived the corresponding influence functions and asymptotic results.
- Designed and conducted simulation studies, grounded in the SIMPLIFY trial design, to evaluate the empirical performance of these causal-inference methods; applied these methods to SIMPLIFY trial analysis.
- Found that augmented inverse propensity weighted (AIPW) outperformed alternatives, achieving the lowest bias, smallest standard errors, and nominal coverage.
- Manuscript in preparation; targeted for JSM 2026 submission.

Benchmarking and fine-tuning genomic language model in ASE prediction 09/2024 - Present
Graduate Research Assistant; Advisor: Professor Pejman Mohammadi, UW Seattle *Seattle, WA*

- Processed RNA-seq data from the Multi-ancestry Analysis of Gene Expression (MAGE) cohort using bioinformatics tools such as bcftools, STAR, and phASER to quantify allele-specific expression (ASE).
- Evaluated the performance of the pre-trained genomic language model **Enformer** and **Borzoi** and biostatistical model **aFc-n** in ASE prediction using genomic information.
- Found that genomic language models pre-trained on reference genomes and genomics data poorly explained individual variance in gene expression and ASE (median Spearman's $\rho = 0.02$–0.05).
- Applied parameter-efficient fine-tuning (using PyTorch Lightning) to **Enformer** and **Borzoi** on RNA-seq expression and ASE data and benchmarked the updated models.
- Fine-tuned Enformer achieved median Spearman's $\rho = 0.192$ on test individuals.

Feature Importance Identifications of Complex Survival Data 10/2022 - 05/2024
Graduate Research Assistant; Advisor: Professor Baiming Zou, UNC Chapel Hill *Chapel Hill, NC*

- Extended the existing method **PermFit**, adapting it for survival data by employing the C-index to compute the permutation feature importance score.
- Enhanced traditional Cox-based deep neural networks (DNN) by incorporating bootstrap aggregating and filtering methods to enhance stability and predictive accuracy (**SurvDNN**).
- Conducted simulation studies of different scenarios. Deduced that under specific simulation scenarios, our proposed method with DNN adeptly identified simulated important variables, exhibiting high power (exceeding 90%) while preserving accurate type I error.
- Tested **SurvDNN** on real-world datasets from breast cancer patients and the Surveillance, Epidemiology, and End Results (SEER) dataset. Established that our method could identify crucial features without compromising prediction accuracy, and **SurvDNN** got superior prediction precision.

Greenspace and Mortality in U.K. 03/2021 - 08/2022
Research Assistant; Advisor: Professor John S. Ji, Tsinghua University *Beijing, CHN*

- Analyzed the association between greenness and mortality using the U.K. Biobank cohort and Cox regression models.
- Employed causal mediation analysis to explore the underlying pathways linking greenness to mortality.
- Concluded that greenness benefits health mainly by reducing air pollution and social isolation, regulating vitamin D and lung function, relieving depression and increasing physical activity.
- This paper was published in *SSM - Population Health*. [Link]

Prevalence of toothache in Chinese adults aged 65 years and above 07/2020 - 03/2021
Research Assistant; Advisor: Professor Jue Liu, Peking University *Beijing, CHN*

- Analyzed data from the Chinese Longitudinal Health Longevity Survey with multivariate modified Poisson regressions with robust error variances to estimate associations between toothache and potential risk factors.
- Found that over 10% of elderly Chinese individuals reported toothaches. Identified several risk factors, including age, gender, socioeconomic status, behavioral patterns, and overall oral health status.
- This paper was published in *Community Dentistry and Oral Epidemiology*. [Link]