Wei Dai

Fairfax, VA

Email: dwei90bd@gmail.com
Mobile: (216) 762 6960
LinkedIn: wdai144
GitHub: wdai0

SUMMARY

New PhD graduate in Statistics (May 2025) with a research focus on feature selection and mixture models; Experienced developer of R and Python packages. Demonstrated experience with several electronic health records (EHR) data; Experienced in machine learning methods, and actively following developments in large language models.

EDUCATION

PhD in Statistics — George Mason University, 2020-2025

- RA on NSF-funded projects; Co-authored 5 papers, contributing to method development and data analysis.
- TA for Statistical Graphics and Data Visualization featuring **R Shiny**, Time Series Analysis, and Regression Models; Best TA Awardee 2024.
- Course training in advanced biostatistics, with a focus on clinical trials and **survival analysis**.

MS in Biostatistics — Case Western Reserve University, 2018-2020

- Maintained 3.90 GPA in statistics and programming coursework; earned **SAS** certification.
- Served on statistical **consulting** team providing statistical support to medical school researchers. Demonstrated experience in communication with collaborators with varying levels of statistical knowledge.
- Collaborated on research projects with Cleveland Clinic Foundation and Veterans Affairs.

PROFESSIONAL EXPERIENCE

Heart Transplant EHR Data Analysis and Allocation Optimization — 2022-2024

Data Scientist

- *Developed predictive models analyzing EHR data to forecast one-year heart transplant survival and performed simulation to optimize donor-recipient matching protocols.*
- Led end-to-end data **pipeline** development, processing sensitive healthcare data on HPC.
- Engineered data preprocessing pipeline, including missing data treatment, and a pilot survival analysis, reducing processing time while preserving data integrity.
- Implemented **machine learning models** (logistic regression, XGBoost, neural networks) on **EHR** data, achieving performance metrics comparable to benchmark publications (different cohorts).
- Briefed on progress through regular meetings and **presentations** to team members from diverse backgrounds, conveying analytical insights, culminating in accepted abstract and presentation at **ISHLT2023** conference.
- paper preprint: *Explainable Machine Learning to Improve Donor-Recipient Matching at Time of Heart Transplant.*

Subcategorizing EHR Diagnosis Codes to Improve Clinical Application of Machine Learning Models — 2021

Data Scientist

- *EHR diagnosis code subcategorization schema developed to improve machine learning model applications in clinical settings. Results provide additional clinical context and temporal information, improving predictive model performance for clinical decision support applications.*
- Identified and validated six diagnosis subcategories; Normalized ICD9/ICD10 codes via UMLS identifiers and integrated them with supporting EHR data.
- Implemented random forest modeling to evaluate subcategorized versus standard diagnosis codes for mortality prediction. Demonstrated 22% improvement in testing AUC using subcategorized model versus standard model.
- Published in *International Journal of Medical Informatics (2021), volume 156, article 104588.*

Atrial Fibrillation (AFib) Recurrence Prediction Enhanced with Biomarkers — 2023

Statistician

- *Statistical analysis conducted on AFib recurrence for a longitudinal study, combining demographic factors and selected biomarkers, resulting in two new biomarker findings.*
- Focused on explaining factors of AFib recurrence, with a secondary task of analyzing AFEQT (quality of life) scores.

- Collaborated with INOVA researchers, progressing beyond t-tests to more sophisticated models including linear mixed effects model and beta regression.
- Presented work to medical doctors and authored in statistical methods section. paper preprint *Novel Biomarkers to Predict Recurrence of Atrial Fibrillation after Catheter Ablation.*

Doctoral Research

- Research focuses on variable selection algorithms and mixture models, validating them through statistical proofs and comprehensive simulation experiments.
- Built **visualization tools** illustrating statistical concepts, improving understanding.
- Implemented **PyTorch-based GPU parallelization** for statistical computing on variable selection, achieving a performance improvement of **20x** over traditional approach.
- Applied algorithm on ovarian cancer omics data; Presented work and poster at JSM 2023 and 2024.
- Created and published open-source packages `subsampwinner` in both R and Python, available on PyPi and **GitHub**.

SKILLS

- Python, R, R shiny
- SAS
- SQL
- Shell, HPC
- Git, GitHub
- Hugging Face, PyTorch
- Generative AI
- AWS

RESEARCH PUBLICATIONS

Articles

[1] Kath M Bogie et al. “Exploring Adipogenic and Myogenic Circulatory Biomarkers of Recurrent Pressure Injury Risk for Persons with Spinal Cord Injury”. In: *J Circ Biomark* 9.1 (Sept. 21, 2020), pp. 1–7. ISSN: 1849-4544, 1849-4544. DOI: 10.33393/jcb.2020.2121.

[2] Dennis Bourbeau et al. “Needs, Priorities, and Attitudes of Individuals with Spinal Cord Injury toward Nerve Stimulation Devices for Bladder and Bowel Function: A Survey”. In: *Spinal Cord* 58.11 (Nov. 2020), pp. 1216–1226. ISSN: 1362-4393, 1476-5624. DOI: 10.1038/s41393-020-00545-w.

[3] Andrew P. Reimer et al. “Patient Factors Associated with Survival after Critical Care Interhospital Transfer”. In: *Front. Disaster Emerg. Med.* 1 (Jan. 8, 2024), p. 1339798. ISSN: 2813-7302. DOI: 10.3389/femer.2023.1339798.

[4] Andrew P. Reimer et al. “Subcategorizing EHR Diagnosis Codes to Improve Clinical Application of Machine Learning Models”. In: *International Journal of Medical Informatics* 156 (Dec. 2021), p. 104588. ISSN: 13865056. DOI: 10.1016/j.ijmedinf.2021.104588.

[5] Jie Xu et al. “Explainable Machine Learning to Improve Donor-Recipient Matching at Time of Heart Transplant”. In: *arXiv preprint* (Mar. 2025).

Packages

[6] Wei Dai and Jiayang Sun. *subsampwinner: A package for feature selection using Subsampling Winner Algorithm.* Python Package Index (PyPI). Version 0.0.8, Accessed: 2025-03-08. Aug. 2024.