Wei Dai

Fairfax, VA

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

SUMMARY

PhD in Statistics with a research experience on feature selection, mixture models, stochastic simulation optimization, computing and real-world data analysis; Experienced R and Python package developer with records & expertise in analyzing electronic health records (EHR) data, applying machine learning methods, and engaging in statistical collaboration with HPC. Actively following and learning advancements in large language models.

EDUCATION

PhD in Statistics — George Mason University, 2020-2025

- Teaching Assistant for Statistical Graphics and Data Visualization featuring **R Shiny**, Time Series Analysis; **Best TA** Recipient 2024.
- Course training in advanced biostatistics, with an emphasis on **clinical trials** and **survival analysis**.
- Dissertation advised by Prof. Jiayang Sun and Prof. Jie Xu: *Statistical Strategies for Contemporary Data Problems: A Generalized Subsampling Winner Algorithm (SWA) for High-Dimensional Variable Selection in Heteroskedastic Data and the Offline Simulation Online Learning (OSOL) Framework for Guided Decision Optimization.*

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

- Earned **SAS** certification; maintained 3.90 GPA in statistics and programming coursework.
- Served on **statistical consulting team** providing statistical support to medical school researchers. Demonstrated experience in communication with collaborators with varying levels of statistical knowledge.

Courses Taken Include

- Survival Analysis; Longitudinal Data Analysis; Advanced Biostatistics; Statistics in Clinical Research.
- Time Series Analysis; Statistical Computing and Optimization.
- Mathematical Statistics; Advanced Stochastic Simulation; Deep Learning (audited).

PROFESSIONAL EXPERIENCE

George Mason University (GMU)

Postdoctoral Research Fellow — **Setp 2025-Present**

- Supported by **GMU-INOVA Joint Postdoc Training Program**.
- Conducting statistical research to real-world challenges.

GMU & INOVA Health System

Biostatistician / Research Assistant — **2022-2024**

- Supported in part by INOVA Health OSRPSD-2759 (PI: Sun) and U19-11-3826 (NIH 1UL1TR003015-1).
- *Developed **predictive models** analyzing OPTN data to forecast one-year heart transplant survival and performed simulation to optimize donor-recipient matching protocols.*
- Participated end-to-end data **pipeline** development, processing sensitive healthcare data on HPC.
- Engineered data pre-processing 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 **OPTN** data, achieving performance metrics comparable to benchmark publications (on different cohorts).
- Briefed on progress through meetings and **presentations** to team members from diverse backgrounds, conveying analytical insights, culminating in accepted abstract and presentation at ***ISHLT2023* conference**.
- Output [5]: *Explainable Machine Learning to Improve Donor-Recipient Matching at Time of Heart Transplant.*

GMU & Cleveland Clinic Critical Care Transport — Jan-June 2021

Data Scientist / Research Assistant

- Supported by NIH R15 1R15NR017792-01 (PI: A Reimer, Co-PI: Sun).
- *EHR diagnosis code subcategorization schema developed to improve **machine learning model applications in clinical settings**. Results provide additional clinical insights 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 at transport. Demonstrated **22% improvement** in testing AUC using subcategorized model versus standard model.
- Output [4] and [3] in *International Journal of Medical Informatics* and in *Front. Disaster Emerg. Med*

GMU & INOVA Health System — Jan-Mar 2023

Biostatistician / Research Assistant

- Supported by INOVA Health U19-11-3826 (PI: Sun).
- *A statistical analysis of a **longitudinal** Atrial Fibrillation (AFib) recurrence study, combining demographic and biomarker data, revealed two new biomarker associations. Additionallay, the study analyzed factors impacting patient quality of life (AFEQT scores).*
- Focused on explaining factors of AFib recurrence, with a secondary task of analyzing AFEQT (quality of life) scores.
- Progressed 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.
- Output: *Novel Biomarkers to Predict Recurrence of Atrial Fibrillation after Catheter Ablation.*

Doctoral Research

- Research focuses on variable selection algorithms, mixture models and stochastic simulation optimization.
- Built **visualization tools** illustrating statistical concepts, improving understanding.
- Implemented **PyTorch based parallelization** for statistical computing on variable selection, achieving a performance improvement of **20x** over traditional approach.
- Presented work and poster at JSM 2023 [14], 2024 [13] and 2025 [12].
- Output: open-source packages `subsampwinner` [6] in both R and Python, available on PyPi and **GitHub**.

PUBLICATIONS AND PRESENTATIONS

ARTICLES

[1] Kath M Bogie, Katelyn Schwartz, Youjin Li, Shengxuan Wang, Wei Dai, and Jiayang Sun. "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, Abby Bolon, Graham Creasey, Wei Dai, Bill Fertig, Jennifer French, Tara Jeji, Anita Kaiser, Roman Kouznetsov, Alexander Rabchevsky, Bruno Gallo Santacruz, Jiayang Sun, Karl B. Thor, Tracey Wheeler, and Jane Wierbicky. "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, Wei Dai, Nicholas K. Schiltz, Jiayang Sun, and Siran M. Koroukian. "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, Wei Dai, Benjamin Smith, Nicholas K. Schiltz, Jiayang Sun, and Siran M. Koroukian. "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, Wei Dai, J. Goldberg, P. Shah, I. Hu, C. Chen, C.R. deFilippi, and J. Sun. "Explainable Machine Learning to Improve Donor-Recipient Matching at Time of Heart Transplant". In: *The Journal of Heart and Lung Transplantation* 42.4, Supplement (2023). ISHLT 43rd Annual Meeting and Scientific Sessions, S22. ISSN: 1053-2498. DOI: https://doi.org/10.1016/j.healun.2023.02.043.

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.

[11] Wei Dai, Jiayang Sun, and Jie Xu. *DACr - An Algorithm for Treating Diverse Missing Values in Large Data, with Application to Heart Transplantation*. Jan. 23, 2026. DOI: 10.64898/2026.01.20.26343799.

MANUSCRIPTS

[7] Wei Dai, Jiayang Sun, and Jie Xu. "An Algorithm for Handling Heterogeneous Missing Data in Real-World Observational Studies, with Application to the SRTR Heart Transplanation Data".

[8] Wei Dai, Jie Xu, and Jiayang Sun. "Connecting Offline Simulation Online Learning (OSOL) through the Lens of Reinforcement Learning".

[9] Wei Dai and Jiayang Sun. "Connecting the Subsampling Winner Algorithm through the Lens of Neural Networks".

[10] Wei Dai, Jiayang Sun, and Jie Xu. "Efficient Multi-fidelity Optimization through Offline Simulation Online Learning (OSOL) Framework".

[11] Wei Dai, Jiayang Sun, and Jie Xu. *DACr - An Algorithm for Treating Diverse Missing Values in Large Data, with Application to Heart Transplantation*. Jan. 23, 2026. DOI: 10.64898/2026.01.20.26343799.

CONFERENCE PRESENTATIONS

[12] Wei Dai. "Efficient Multi-fidelity Optimization through Offline Simulation Online Learning (OSOL) Framework". Joint Statistical Meetings 2025 (Nashville, TN). Apr. 8, 2025.

[13] Wei Dai. "Subsampling Winner Algorithm 2.0: Feature Selection in High-Dimensional Heteroskedastic Data and More". Joint Statistical Meetings 2024 (Portland, OR). June 8, 2024.

[14] Jiayang Sun and Wei Dai. "Subsampling Winner Algorithm for Feature Selection". Joint Statistical Meetings 2023 (Toronto, ON, Canada). Aug. 8, 2023.

[15] Wei Dai. "Offline Simulation + Online Learning with E-Learning: Decision Support Using High and Low Fidelity Data". The 9th International Workshop in Sequential Methodologies (Washington, DC, USA). June 2026.