Zhengyang Fei

437-580-0363 johnyzyfei.fei@outlook.com linkedin.com/in/zhengyang-fei github.com/johnyzyfei Website

Education

University of Toronto *Toronto, Canada*
M.Sc. Biostatistics, Dalla Lana School of Public Health **Nov 2025**
Emphasis in AI and Data Science
Relevant Coursework: Categorical Data Analysis, Survival Analysis, Machine Learning, Applied Bayesian Methods

University of Toronto *Toronto, Canada*
HB.Sc. Applied Statistics and Mathematics **Jun 2024**
Relevant Coursework: Data Analysis, Experimental Design, Statistical Learning, Applied Time Series Analysis

Technical Skills

Statistical and Programming Tools: R, Python, SAS, SQL, JAGS
Productivity and Documentation:: Git, Microsoft Office Suite, MiKTeX, Adobe Premiere Pro, Power BI

Experience

Institute for Better Health - Trillium Health Partners **May 2025 - Aug 2025**
Data Analyst - Student Researcher (Supervisor: Dr. Simona Minotti) *Mississauga, Ontario*

- Created research-ready datasets by cleaning and transforming high-dimensional patient experience data; applied data quality checks to support population health surveillance and equity analyses.
- Documented data cleaning steps and maintained data dictionaries to ensure reproducibility and team alignment.
- Researched methods for extracting themes from free-text using text mining and natural language processing (NLP).
- Used R for analysis, including the I-MAIHDA multilevel model for identifying disparities in patient-reported experience across intersecting social strata, informing hospital equity strategies.
- Collaborated with a team of multidisciplinary researchers to ensure data quality and consistency.
- Delivered actionable insights to hospital leadership through formal reports and presentations, influencing quality improvement priorities; contributed to a manuscript under preparation.

University of Toronto - MiDATA Lab **Nov 2024 - Present**
Data Analyst - Researcher/Biostatistician *Toronto, Ontario*

- Conducted biomarker analysis in R to evaluate haptoglobin against established inflammatory biomarkers using generalized linear mixed models (GLMM) in longitudinal juvenile arthritis study.
- Supported statistical analysis in Python using pandas, statsmodels, and scikit-learn to classify treatment response based on ADC-derived imaging biomarkers via penalized logistic regression and ROC analysis in a cancer cohort.
- Conducted data cleaning, profiling, and visualization of clinical datasets; supported methodological research through literature review and implementation of appropriate statistical techniques for clinical studies.
- Interpreted analytical results and collaborated with a multidisciplinary team to ensure data quality, draft summary reports, and contribute to manuscript preparation for journal submission.

University of Toronto **Sep 2022 - Present**
Teaching Assistant (Link) *Toronto, Ontario*

- Fulfilled core TA responsibilities across multiple undergraduate courses—including differential, integral and multivariable calculus, linear algebra, probability and statistics, statistical modeling, and data analysis—by leading tutorials, grading assessments, and hosting office hours.
- Served as Head TA for the Statistical Modelling course, creating lecture videos, co-developing course materials with the coordinator, and designing marking schemes to support consistent evaluation.

Projects and Competitions

Atrial Fibrillation Prediction Modeling Using ECG and EHR Data (Link) **2025**

- Performed missing data imputation using the MICE package with predictive mean matching.
- Conducted feature selection and model tuning using ECG and EHR data.
- Trained and optimized survival gradient boosting and random forest models to predict new-onset atrial fibrillation.
- Presented project findings as a poster at the 2025 Statistical Society of Canada (SSC) Student Research Competition.

Multiple Imputation Deletion R Package (Link) **2024**

- Developed an R package (DIMP) that addresses the issue of Multiple Imputation then Deletion by removing imputed outcomes, enabling users to perform analyses and build models using observed outcomes alongside imputed covariates.