Doctoral Dissertation Defense Announcement

“Machine learning approaches for single-cell RNA-seq study design and composite endpoint data”

Yu Wang
Candidate for Doctor of Philosophy
Biostatistics
School of Graduate Studies
Medical College of Wisconsin

Committee in Charge:
Chien-Wei Lin PhD (Chair, Advisor)
Kwang Woo Ahn, PhD (Co-advisor)
Soyoung Kim, PhD
Paul Auer, PhD
Yi-Guang Chen, PhD

Date: Tuesday, June 18, 2024
Time: 10:00 AM (CST)
Defense Location: M2050/M2070

https://mcw-edu.zoom.us/j/96167811848?pwd=RloyPzansVsaPLw71ArpX5CsGWJfal.1
Meeting ID: 961 6781 1848 Passcode: Aa1yCuQE
Graduate Studies:
Mathematical Statistics I
Mathematical Statistics II
Research Seminar
Biostatistical Computing
Statistical Models & Methods I
Statistical Model & Methods II
Linear Models I
Introduction to Epidemiology
Theory of Statistical Consult
Applied Survival Analysis
Theory of Survival Analysis
Statistical Consulting
Reading and Research
Intro to Bayesian Analysis
Intro to Translational Bioinformatics
Ethics & Integrity in Science
Design & Analysis Clin Trials
Statistical Genetics
Introduction to Statistical Machine Learning
Advanced Statistical Computing
Advanced Bayesian Analysis
Advanced Statistics I
Problem Solving with Object-Oriented Programming
Research Ethics Discussion Series
Doctoral Dissertation
Dissertation
“Machine learning approaches for single-cell RNA-seq study design and composite endpoint data”

Machine learning also plays a vital role in life sciences from the molecular to the individual. With advancements in technology in molecular biology, Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful technology for characterizing transcriptomic profiles at single-cell resolution. At the stage of planning such experiments, it is crucial to consider both the number of cells and sequencing depth. While existing literature address this need, they are primarily simulation-based methods and rely on Unique Molecular Identifier (UMI) matrix which has critical limitations such as ignoring the actual number of reads needed. Here we propose the first FastQ-based study design framework for scRNA-seq data, named “FastQDesign,” which leverages FastQ files from publicly available scRNA-seq datasets as references and suggests an optimal design within a given budget. We demonstrate the feasibility of this framework through simulations and applications to three real-world scRNA-seq datasets. Our study shows the need to have an appropriate design to study the underlying heterogeneous cell population. A high-efficiency software suite is provided and can be found at https://github.com/yuw444/FastQDesign.

On the other hand, medical research often involves the study of composite endpoints that combine multiple clinical events to assess the efficacy of treatments. When constructing composite endpoints, it is a common practice to analyze the time to the first event. However, this approach overlooks outcomes that occur after the first event, resulting in information loss. Furthermore, the terminal event can not only be of interest, but also a competing risk for other types of outcomes. While regression models exist to analyze all types of such outcomes, not just the first event, and properly address the terminal event, they do not account for potential nonlinear covariate effects on the logarithm of the rate function. To address these important issues, we introduce random forests for composite endpoints (Rforce) consisting of non-fatal composite events and terminal events. Rforce utilizes generalized estimating equations to build trees and handles the dependent censoring due to the terminal events with the concept of pseudo-at-risk-duration. Simulation studies and real data analysis are conducted to demonstrate the performance of Rforce(https://github.com/yuw444/Rforce).
Yu Wang, Ph.D. Candidate
Milwaukee, WI · (417)-450-0271 · yuw926@icloud.com · · ·

Education

Ph.D. Candidate in Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA 08/2019-07/2024
Certificate in Business Analyst, Missouri State University, Springfield, MO, USA 08/2018-05/2019
M.S. in Mathematics, Missouri State University, Springfield, MO, USA 08/2016-05/2018
B.S. in Statistics, Henan University of Economics and Law, Zhengzhou, China 09/2012-06/2016

Ph.D. Dissertation

Title: Machine learning approaches for single-cell RNA-seq study design and composite endpoint data

- **Aim 1**: FastQDesign: A Realistic FASTQ-based Framework for ScRNA-seq Study Design Issues. [Yu Wang et al.] To Be Submitted. It is a bioinformatics tool that addresses the power analysis of the scRNA experiment design by using publicly available fastq reference datasets.
- **Aim 2**: Rforce: Random forests for composite endpoints. [Yu Wang et al.] To Be Submitted. It aims to estimate the mean event rates of composite endpoints given the patient’s characteristics.
- **Aim 3**: Package Development. Develop the packages for both aims 1 and 2.

Advisors: Prof. Chien-Wei Lin, Kwang Woo Ahn

Experience

Research Assistant 12/2022 - present
Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI

- scRNA-seq trajectory comparison between the condition, one paper is under preparation.
- Toxicity monitoring rule for a two-cohort Phase II clinical trial with bivariate Beta prior, one paper with an R package is under preparation. [Yu Wang and Szabo] To Be Submitted.
- Single-cell RNA-sequencing (scRNA-seq) data analysis in investigating Type 1 diabetes with the mouse model. [Ciecko et al.] 2023
- Provided the comparative data analyses for before and after breast cancer therapy. [Chaudhary et al.] 2023.
- Investigated the creatine kinase level among the different etiologies for rhabdomyolysis patients.
- Performed the survival data analysis for primary pediatric lung malignancies. [Smith et al.] 2021.

Teaching Assistant 08/2016 - 05/2018
Mathematics Department, Missouri State University, Springfield, MO

- Differential Equations with Dr. Shouchuan Hu
- Statistical Methods with Dr. Yingcai Su
- Intermediate Algebra with Instructor Jennifer Pursley

Invited Presentation


Yu Wang, Soyong Kim, Chien-Wei Lin, Kwang Woo Ahn. Rforce: Random Forest for Composite Endpoints (MCW Graduate School Travel Award). ENAR, Baltimore, MD, 03/2024

Yu Wang, Chien-Wei Lin. FastQDesign: A Realistic FASTQ-Based Framework for ScRNA-Seq Study Design Issues (MCW Graduate School Travel Award). ENAR, Nashville, TN, 03/2023

Yu Wang, Aniko Szabo. Toxicity monitoring rule for a two-cohort Phase II clinical trial with bivariate Beta prior (MCW Graduate School Research Poster Award), Milwaukee, WI, 01/2021

Publications


† indicates co-first author

## Awards

- ASA Statistical Computing and Graphics Student Paper Competition Award 01/2024
- Travel Award, Medical College of Wisconsin 05/2023
- Research Poster Award, Medical College of Wisconsin 03/2021
- Medical College of Wisconsin Fellowship 08/2019 - 12/2020
- Graduate Assistantship, Missouri State University 08/2016 - 05/2018
- Dean’s List, Missouri State University 12/2015 - 05/2019
- Public Speaking Competition Award, Henan Uni. of Economics & Law 10/2014
- Scholarship Prize, Henan Uni. of Economics & Law 09/2014
- Outstanding Volunteer, Henan Uni. of Economics & Law 05/2014
- Scholarship Prize, Henan Uni. of Economics & Law 09/2013

## Mentor Experience

<table>
<thead>
<tr>
<th>Mentor</th>
<th>Coach</th>
<th>Scholar</th>
</tr>
</thead>
</table>

## Academic Society

<table>
<thead>
<tr>
<th>Role</th>
<th>Institution/Region</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer</td>
<td>Special Issue on Biostatistics in Animal Research in Laboratory Animals</td>
<td>01/2024 - present</td>
</tr>
<tr>
<td>Member</td>
<td>American Statistical Association</td>
<td>09/2022 - present</td>
</tr>
<tr>
<td>Member</td>
<td>International Biometric Society Eastern North American Region (Member)</td>
<td>09/2022 - present</td>
</tr>
<tr>
<td>Member</td>
<td>RISC-V International</td>
<td>05/2023 - present</td>
</tr>
<tr>
<td>Member</td>
<td>International Chinese Statistical Association</td>
<td>03/2023 - present</td>
</tr>
</tbody>
</table>

## Technical Skill

<table>
<thead>
<tr>
<th>Languages</th>
<th>Developer Tools</th>
<th>IDEs</th>
<th>Platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>English; Mandarin; Spanish(A1)</td>
<td>R; C; Python; Git; Java; SQL; SAS; Conda; Docker; Singularity</td>
<td>RStudio; VS Code</td>
<td>High-performance Computing Cluster; Linux; Mac; Windows</td>
</tr>
</tbody>
</table>