INTERNATIONAL CHINESE STATISTICAL ASSOCIATION

泛華統計協會

Applied Statistics Symposium

Data-Driven Decision Making: Unleashing the Power of Statistics

June 16–19, 2024

Nashville, Tennessee

Hosted by

VANDERBILT UNIVERSITY MEDICAL CENTER

Department of Biostatistics

Photo: View of Nashville from the roof of Loews Vanderbilt. © Vanderbilt University

#ICSAnashville

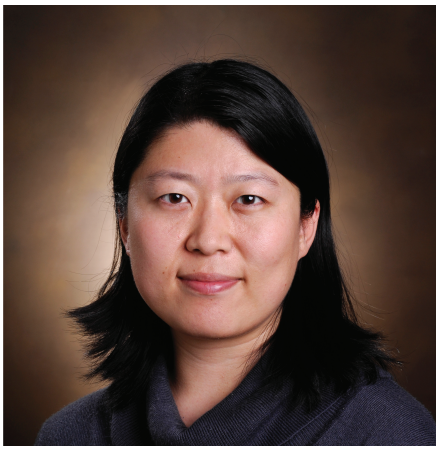
Contents

Click a section title to go to it.

Welcome to the 2024 ICSA Applied Statistics Symposium

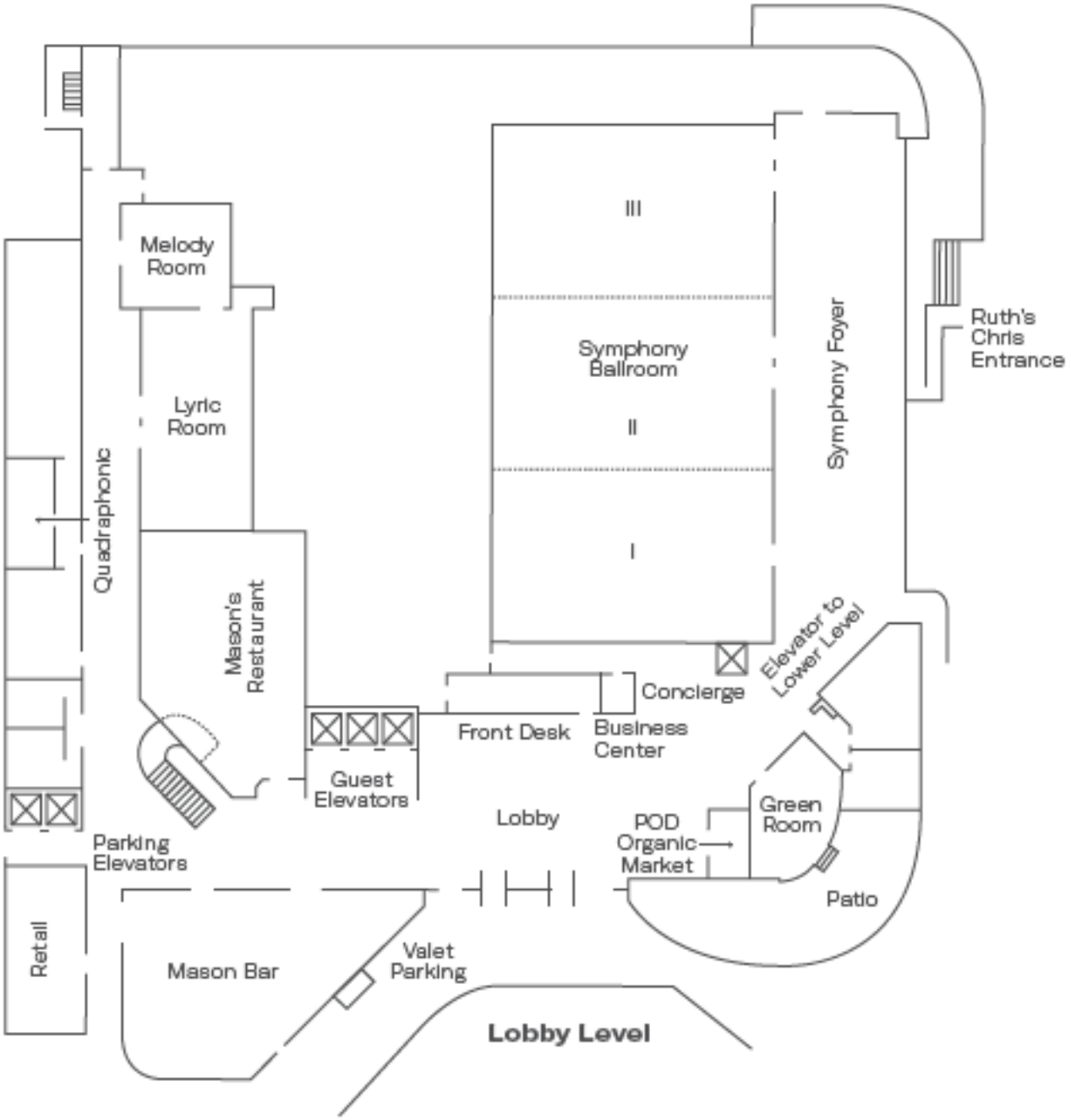
As executive co-chairs of the 2024 ICSA Applied Statistics Symposium organizing committees, it is our pleasure to welcome you to Nashville, Tennessee, USA, home of Vanderbilt University Medical Center and a global hub for numerous educational institutions, healthcare corporations, and technological enterprises, as well as a destination for music lovers, sports fans, conference attendees like yourself, and more. We thank the ICSA site selection committee for accepting our bid to host the 33rd edition of the symposium, and all the sponsors and other contributors of funding, time, and resources toward the success of this year's gathering.

We are tremendously excited about the courses, panels, posters, and talks that are being presented here, as well as other opportunities to connect with statisticians from all over in discussing "Data-Driven Decision Making: Unleashing the Power of Statistics." With so much going on in our communities and the wider world, these occasions for sharing our expertise, learning from one another, and amplifying our discoveries and findings are crucial to shaping a better future for us all.



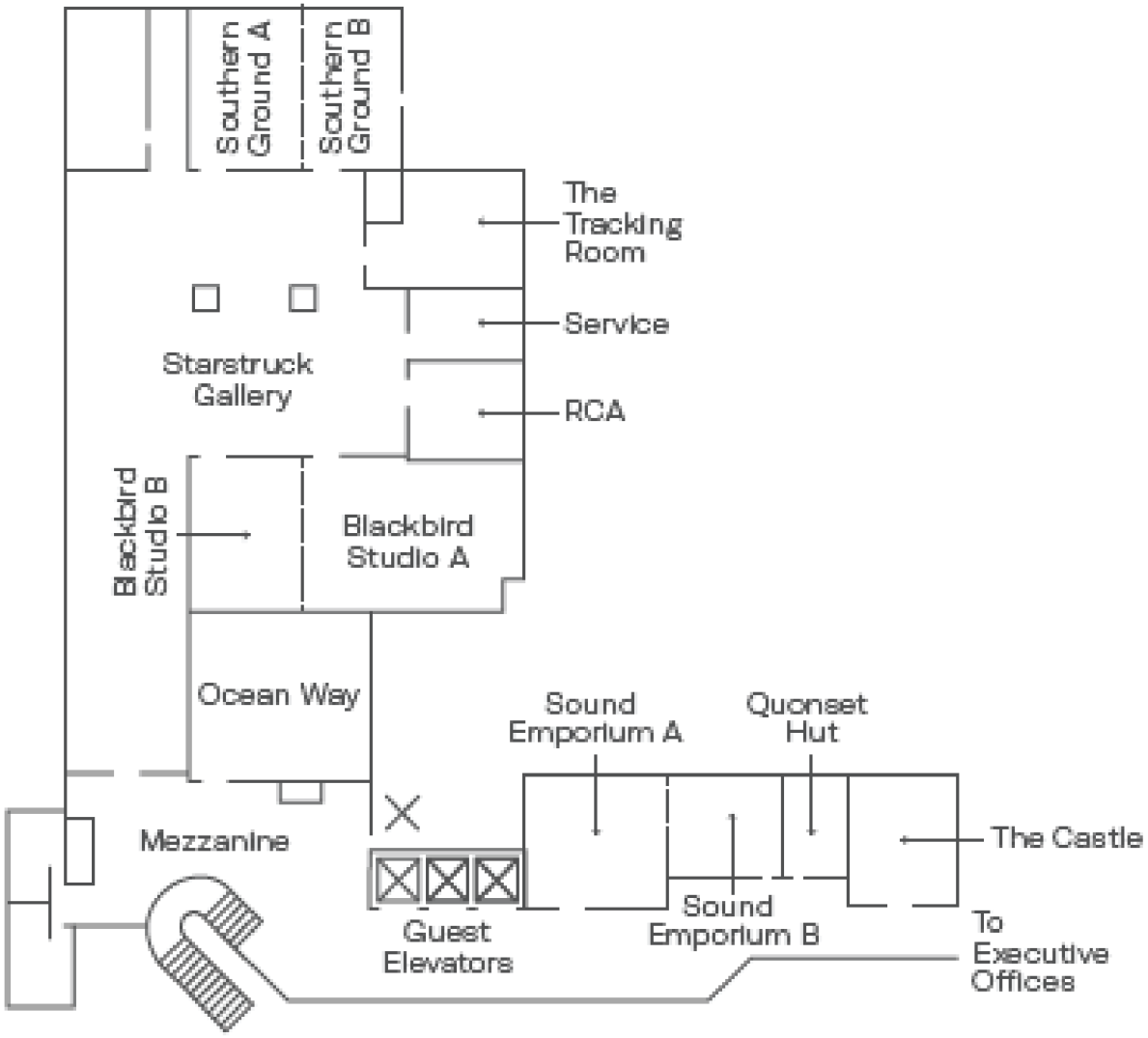
Dandan Liu, PhD
Associate Professor of Biostatistics
Executive Director, Vanderbilt Biostatistics Data Coordinating Center (VBDCC)
Director, Vanderbilt Institute for Clinical and Translational Research (VICTR) Methods Program
Vanderbilt University Medical Center

Qingxia "Cindy" Chen, PhD
Professor of Biostatistics, Biomedical Informatics, and Ophthalmology & Visual Sciences
Vice Chair of Education and Director of Post-Graduate Studies and Distance Learning, Biostatistics
Director, Executive Data Science Program, Departments of Biostatistics and Biomedical Informatics
Vanderbilt University Medical Center

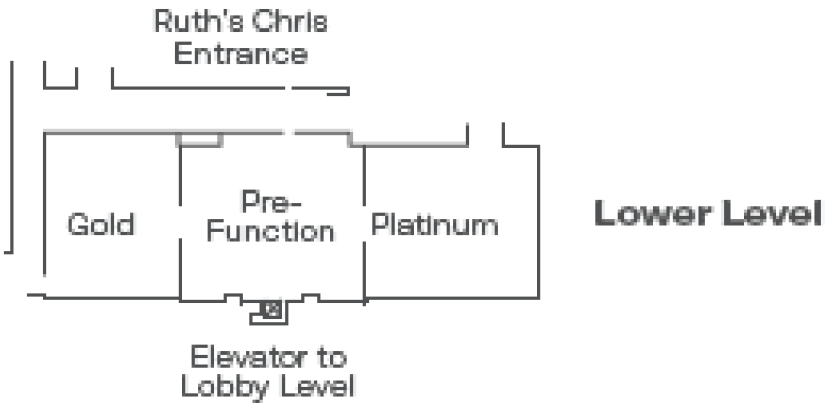


Loews Nashville Hotel
at Vanderbilt Plaza

2100 West End Avenue
Nashville, TN 37203
(615) 320-1700



Mezzanine Level



Lower Level

Acknowledgments

谢谢

We thank all our sponsors for supporting the 2024 ICSA Applied Statistics Symposium.

U.S. National Science Foundation
WHERE DISCOVERIES BEGIN

The symposium is supported in part by the U.S. National Science Foundation under Award No. 2410953.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Gold Sponsors

VANDERBILT UNIVERSITY MEDICAL CENTER
Department of Biostatistics

abbvie

Boehringer Ingelheim

THE LOTUS GROUP

Pfizer

#ICSAnashville

Silver Sponsors

AMGEN

BeiGene

Bristol Myers Squibb™

EVEREST
CLINICAL RESEARCH

GILEAD
Creating Possible

Lilly

mathematics
an Open Access Journal by MDPI

MERCK

sanofi

Bronze Sponsors

Please visit the symposium Career Services webpage (symposium2024.icsa.org/career-services) to learn about job opportunities with some of these sponsors. Several recruiters will be conducting on-site interviews and info sessions in the Tracking Room (Mezzanine Level) during the symposium.

Visit the sponsor exhibition tables in the Symphony Entrance Foyer (Lobby Level) for information, giveaways, and more!

VANDERBILT UNIVERSITY MEDICAL CENTER | Department of Biostatistics

Center for Quantitative Sciences

2024 Summer Institute

Four great courses—one virtual, three in-person.
Pick one or more to expand your research and analytical skill sets.

All of Us RESEARCH PROGRAM | The Future of Health Begins With You

ONLINE
Introduction to the *All of Us* Research Program
July 22–26, 1–4 p.m. Central Time via Zoom
Course directors: Paul Harris and Cindy Chen

ON CAMPUS
at 2525 West End Avenue (parking and food included)

Big Data in Biomedical Research
July 22–26, 9 a.m.–noon
Instructors: Qi Liu and Yu Shyr

Regression and Modeling in R
July 29–August 2, 9 a.m.–noon
Instructor: Gustavo Amorim

Introduction to Causal Inference
July 29–August 2, 1–4 p.m.
Instructor: Andrew Spieker

Scan this code or visit
vumc.org/cqs/cqs-summer-institute
to learn more and to register.

Find out more
about Vanderbilt Biostatistics at
vumc.org/biostatistics

Interested in our graduate training program?
Go to vanderbilt.edu/biostatistics-graduate for details.

Why Choose The Lotus Group?

The Lotus Group LLC is a certified minority women-owned (MWBE) business delivering biometrics recruiting services nationally. In 2020, we were recognized as one of the 50 fastest growing women-owned businesses by the Women Presidents' Organization, sponsored by American Express.

Our team members are located throughout the country including California, Florida, Massachusetts, and New Jersey; many with over 20 years of staffing experience in the pharmaceutical industry.

Our consultative approach and knowledge of the market ensure that your career is in good hands. We see ourselves as your "career advocates" as we provide you with everything you need to successfully navigate your job search and advance your career.

Step on to The Lotus Group bridge – connecting great biometrics candidates and companies – and start your journey to success!

Key Functional Areas We Support

- Biostatistics
- Statistical Programming
- Clinical Programming
- Data Management
- Data Science
- HEOR/RWE
- Epidemiology
- Pharmacology
- Drug Safety
- Clinical Development
- Clinical Operations

How to Contact & Follow Us

- Check out our website: https://tlgcareers.com/
- Follow us on LinkedIn - scan the QR Code
- Send resumes or inquiries to info@tlgcareers.com

Meet Sanofi.
Life might change thanks to you.

Sanofi is an innovative global healthcare company, driven by one purpose: we chase the miracles of science to improve people's lives.

We are dedicated to transforming the practice of medicine by working to turn the impossible into the possible. We provide potentially life-changing treatment options and life-saving vaccine protection to millions of people globally, while putting sustainability and social responsibility at the center of our ambitions.

Our Biostatisticians at Sanofi play a key role in the innovation journey by investing in the development of all our team members, offering compelling and exciting career opportunities that value diversity of thought and abilities. We have a shared a commitment to bring statistical innovation and rigor to our competitive portfolio, ranging over immunology, oncology, neurology, rare disease, and rare blood disorders. Adeptly pioneering the digital AI frontier, we are building upon and expanding our core expertise and leadership in data and statistical science in our quest to bring transformational therapies to patients.

Follow us

sanofi

To learn more about Sanofi, please visit Sanofi Careers: jobs.sanofi.com/en

mathematics
an Open Access Journal by MDPI

mdpi.com

AMGEN

Amgen harnesses the best of biology and technology to fight the world's toughest diseases, and make people's lives easier, fuller and longer. We helped establish the biotechnology industry, and we remain on the cutting-edge of innovation, using technology and human genetic data to push beyond what's known today.

amgen.com

LILLY FOR BETTER SCIENCE

We live in an amazing era for medicine. At Lilly, we use groundbreaking science to meet unmet medical need in the areas of diabetes, oncology, immunology, neurodegeneration and pain. Our determination is to advance the best science of the day in an effort to make life better for people around the world.

Learn more about how we're using science to make life better at lilly.com.

2020 CA Approved for External Use PRINTED IN USA ©2020, Eli Lilly and Company. ALL RIGHTS RESERVED.

Lilly

ecrscorp.com

EVEREST

CLINICAL RESEARCH

We are a full service CRO
built from a statistical and data management center of excellence.

We provide a broad range of expertise-based clinical research services to worldwide pharmaceutical, biotechnology, and medical device industries.

abbvie

Meeting the needs of patients—and the needs of our time.

abbvie.com

Images © AbbVie Inc. All rights reserved.

Boehringer Ingelheim

Save the date!

Join us at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, for our next Transforming Science Day: February 25, 2025.

Boehringer Ingelheim has 176 affiliates worldwide, with over 52,000 employees.
Our locations include Argentina, Brazil, China, France, Japan, Singapore, Taiwan, and more.

boehringer-ingelheim.com

For 175 years, Pfizer has been a trusted partner in healthcare – discovering, developing, and delivering medical breakthroughs to prevent, treat, and cure some of the world's most vexing conditions and diseases.

pfizer.com

BeiGene Corporate Overview

Beigene is a global oncology company that was built differently to deliver innovative medicines faster, more equitably and affordably around the world.

Our founding belief is that there is a better way to bring innovative treatments to patients around the world. We are an oncology powerhouse with a deep, diverse pipeline fueled by one of the industry's largest and most productive research teams.

Our two foundational medicines, BTK inhibitor BRUKINSA® (zanubrutinib) and PD-1 inhibitor TEVIMBRA® (tislelizumab), demonstrate the strength of our science and our mission to improve treatment outcomes for patients.

Today, **more than 10,000** colleagues operate in more than **40 markets** across **five continents**. **More than 1 million** patients have been treated with our medicines, reflecting our expansive global reach and deep commitment to access.

Facts at a Glance

10k+ Colleagues globally in over **40** offices on **5** continents

$2.5B Annual product revenue **$3.2B** cash balance*

1M+ Patients treated with our medicines

30+ Assets in clinical and commercial stages

3.7k+ Global commercial team members

In-house manufacturing including U.S. expansion

1.1k+ Oncology research team

40+ Phase 3 or potentially registration enabling trials

~20 Industry collaborations

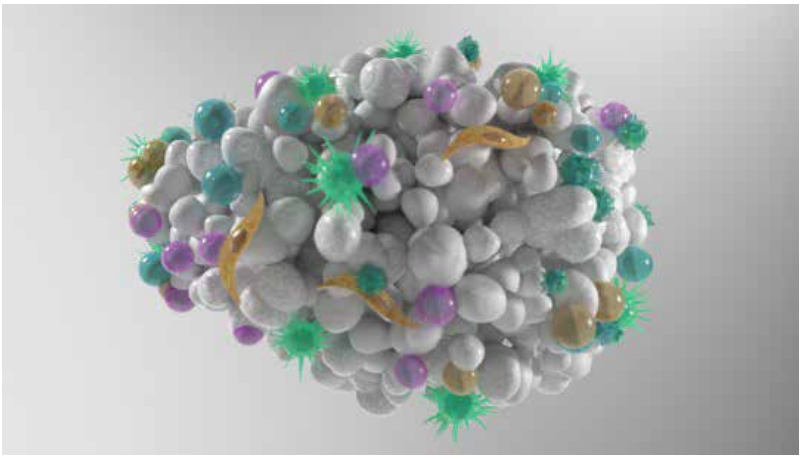
*As of February 26, 2024

For more information visit:

BeiGene.com | @BeiGeneGlobal | Linkedin.com/Company/BeiGene | BeiGene Global

BeiGene

Bristol Myers Squibb™



Building a better future

bms.com

Powered by research.

For more than a century, we've been at the forefront of research, bringing forward medicines, vaccines and innovative health solutions for some of the world's most challenging diseases.

We use the power of leading-edge science to save and improve lives around the world.

Research is in our DNA.

MERCK

Copyright © 2022 Merck & Co., Inc., Rahway, NJ, USA and its affiliates. All rights reserved. US-NON-11581 09/22

Your satisfaction is our mission
Your success is our passion

LLX Solutions, LLC

Providing Top-quality Statistical, Programming, Data Management and Consulting Services to Biotech/Pharmaceutical and Medical Device Industries

VERTEX®

At Vertex, we invest
in scientific innovation
to create transformative medicines
for people with serious diseases.

vrtx.com

REGENERON
SCIENCE TO MEDICINE®

PUSH THE BOUNDS OF SCIENCE

MAKE LIFE-CHANGING MEDICINES

regeneron.com

VANDERBILT
Data Science

vanderbilt.edu/datascience

ICSA Leadership

Executives

President: Xun Chen
President-Elect: Hongyu Zhao
Treasurer: Rui Feng
Past-President: Gang Li
Executive Director: Jun Zhao
Office Manager: Grace Ying Li

Board of Directors

Gang Li, Annie Qu, Lan Wang, Hao (Helen) Zhang, Xingqiu Zhao, Ming Tan, Huazhen Lin, Min Zhang, Li Wang, Yanping Wang, Jialiang Li, George Tseng, Kun Chen, Song Yang, Jianchang Lin

Program Committee

Xinping Cui (chair), Hulin Wu, Jianguo (Tony) Sun, Yingwen Dong, Shuangge Ma, Gongjun Xu, Jian Kang, Dandan Liu, Qingxia Chen, Qiqi Deng, Xiaojing Wang, Yichuan Zhao, Xin-Yuan Song, Ying Zhang, Ming-Chung Chang, Lihui Zhao

Awards Committee

Zhigang Li (chair), Chunming Zhang, Wei Wu, Lu Tian, Yong Chen, Xuezhou Mao, Bo Huang, Charles Ma, Jiayang Sun, Hong Tian, Huilin Li, Gen Li, Kai Yang

Nominating and Election Committee

Yichuan Zhao (chair) Wenqing He, Hongjian Zhu,Zhigang Li, Tiejun Tong, Li-Shan Huang, Wei Zhang, Jin-Ting Zhang

Special Lecture Committee
Ming Tony Tan (chair), Hongzhe Lee, Aiyi Liu, Gang Li, Jianguo Sun, Xiaonan Xue

Publication Committee
Runze Li (chair), Linda Zhao, Hongkai Ji, Jianguo (Tony) Sun, Yi-Hau Chen, John Stufken, Huixia Judy Wang, Ding-Geng (Din) Chen, Chixiang Chen, Jun Zhao, Grace Ying Li, Sheng Luo, Ying Ding

Membership Committee
Zhigen Zhao (chair), Xun Chen, Fei Huang, Tiejun Tong, Tu Xu, Wei Zhang, Anru Zhang

IT Committee
Chengsheng Jiang (chair)

Archive Committee
Naitee Ting (chair), Xin (Henry) Zhang, Rui Miao, Xin Tian, Jun Yan

Finance Committee
Rui Feng (chair), Xin He, Rochelle Fu

Financial Advisory Committee
Fang Chen (chair), Nianjun Liu, Xiangqin Cui, Rochelle Fu, Yuan Jiang, Hongliang Shi

Lingzi Lu Award Committee (ASA/ICSA)
Ivan Chan (chair), Kelly Zou, Shelly Hurwitz, Laura J. Meyerson

Representatives to JSM Program Committee
Jianguo (Tony) Sun (2023), Yingwen Dong (2024), Shuangge Ma (2025)

Outreach and Engagement Committee
Jin Zhou (chair), Qing Yang (chair), Jun Zhao, Weining Shen, Gang Li, Chengsheng Jiang, Grace Ying Li, Chixiang Chen

Constitution Committee
Hongzhe Lee (chair), Rochelle Fu, Jun Zhao, Zhezhen Jin, Ying Lu, Jianguo Sun, Chengsheng Jiang, Yichuan Zhao, Jiayang Sun

Chapter Chairs
ICSA-Canada: Joan Hu
ICSA-Midwest: Xiaohong Huang
ICSA-Taiwan: Henry Horng-Shing Lu

2024 Symposium Committees

Executive Committee

Dandan Liu, Vanderbilt University Medical Center – symposium co-chair
Qingxia "Cindy" Chen, Vanderbilt University Medical Center – symposium co-chair
Gang Li, University of California at Los Angeles – ICSA president (2023)
Jun Zhao, Antengene – ICSA executive director
Xun Chen, Sanofi – ICSA president (2024)
Rui Feng, University of Pennsylvania – ICSA treasurer
Chengsheng Jiang, University of Maryland, College Park – ICSA IT
Xinping Cui, University of California, Riverside – ICSA program committee chair

Scientific Program Committee

Dandan Liu, Vanderbilt University Medical Center – co-chair
Qingxia "Cindy" Chen, Vanderbilt University Medical Center – co-chair
Erik Bloomquist, Merck
Yong Chen, University of Pennsylvania
Zhen Chen, NIH/NICHD/DIPHR
Yu Cheng, University of Pittsburgh
Gaohong Dong, BeiGene
Yingwen Dong, Sanofi
Haoda Fu, Lilly
Yijuan Hu, Emory University
Chao Huang, Florida State University
Hongkai Ji, Johns Hopkins University
Chun Li, University of Southern California
Gang Li, EISAI
Yimei Li, University of Pennsylvania
Jianchang Lin, Takeda
Bin Nan, University of California, Irvine
Jing Ning, MD Anderson
Tony (Jianguo) Sun, University of Missouri
Sijian Wang, Rutgers University
Xiaojing Wang, University of Connecticut
Yanping Wang, Lilly
Ying Wei, Columbia University
Gongjun Xu, University of Michigan
Zhenzhen Xu, US Food and Drug Administration
Shu Yang, North Carolina State University
Min Zhang, University of Michigan
Song Zhang, UT Southwestern Medical Center
Yichuan Zhao, Georgia State University
Yingqi Zhao, Fred Hutch
Yize Zhao, Yale University
Donglin Zeng, University of Michigan
Jin Zhou, UCLA
Fei Zou, University of North Carolina at Chapel Hill

Poster Session Committee

Fei Ye, Vanderbilt University Medical Center - chair

Judges

Anjun Cao, Bayer U.S. LLC
Sheau-Chiann Chen, Vanderbilt University Medical Center
Chongzhi Di, Fred Hutchinson Cancer Center
Wenyun Gao, University of North Carolina at Charlotte
Lin Hou, Tsinghua University
Mi-Ok Kim, University of California San Francisco
Liang Li, MD Anderson Cancer Center
Jeen Jing-ou Liu, Regeneron
Rachael Liu, Takeda
Yushi Liu, Eli Lilly and Company
Ting Fung Ma, University of South Carolina
Kelly Street, University of Southern California
Shikun Wang, Columbia University
Chun Yip Yau, Chinese University of Hong Kong
Lin Zhang, University of Minnesota
Xuwen Zhu, University of Alabama

Poster Session Committee
Fei Ye, Vanderbilt University Medical Center - chair

Student Paper Competition Committee
Ran Tao, Vanderbilt University Medical Center - co-chair
Siyuan Ma, Vanderbilt University Medical Center - co-chair

Reviewers
Ai Ni, The Ohio State University
Bingxin Zhao, University of Pennsylvania
Chang Su, Emory University
Didong Li, University of North Carolina at Chapel Hill
Fangzheng Xie, Indianna University Bloomington
Fei Gao, Fred Hutchinson Cancer Center
Gen Li, University of Michigan
Haoyu Zhang, National Cancer Institute
Hu Guanyu, University of Texas Health Science Center at Houston
Jichun Xie, Duke University
Jin Zhou, University of California, Los Angeles
Lihua Lei, Stanford University
Linxi Liu, University of Pittsburgh
Lu Mao, University of Wisconsin-Madison
Lu Xia, Michigan State University
Ming Zhou, Bristol Meyers Squibb
Nianqiao Ju, Purdue University
Qi Zheng, University of Louisville
Qihuang Zhang, McGill University
Qingning Zhou, University of North Carolina at Charlotte
Revathi Ananthakrishnan, Bristol Meyers Squibb
Rong Ma, Harvard University
Ruoqing Zhu, University of Illinois Urbana-Champaign
Shaoyang Ning, Williams College
Siddhesh Kulkarni, Bristol Meyers Squibb
Siyu Heng, New York University
Wodan Ling, Cornell University
Xiao Wu, Columbia University
Xinran Li, University of Chicago
Xuan Bi, University of Minnesota
Ying Ma , Brown University
Yuan Huang, Yale University

Short Course Committee
Qingxia “Cindy” Chen, Vanderbilt University Medical Center – co-chair
Jinyuan Liu, Vanderbilt University Medical Center – co-chair

Financial/Business Committee
Janey Wang, Vanderbilt University Medical Center - chair
Rui Feng, University of Pennsylvania

Fundraising Committee
Jane Zhang, AbbVie – chair
Xiaoming Xu, AbbVie
Tu Xu, Novo Nordisk
Zeqing Lu, Eli Lilly
Meng Cao, Novartis
Jiarui Chi, Sanofi
Xin Chen, AstraZeneca
Yanan Huo, Gilead
Pranab Mitra, Takeda

Local Committee
Dandan Liu, Vanderbilt University Medical Center – co-chair
Panpan Zhang, Vanderbilt University Medical Center – co-chair
Margaret Cullum, Vanderbilt University Medical Center
Jena Altstatt, Vanderbilt University Medical Center

Program Book and Website Committee
Yaomin Xu, Vanderbilt University Medical Center – co-chair
Shawn Garbett, Vanderbilt University Medical Center – co-chair
Peg Duthie, Vanderbilt University Medical Center
Siwei Zhang, Vanderbilt University Medical Center

Volunteers
Kun Bai
Eric Yongxin Chen
Huiding Eric Chen
Andrew Gan
Sophie Gao
Xiaoming Gao
Yue Gao
Kaidi Kang
Eric Koplin
Xiangyu Ji
Lisa Levoir
David Liu
Kaixing Liu
Sydney Louit
Tianyi Sun
Shengxin Tu
Hao Wu
Shiying Xiao
Bailu Lucy Yan
Lydia Yao
Chih-Ting Yang
Haoyang Yi
Siwei Zhang

Sunday, June 16, 2024

8:00 am – 5:00 pm	Registration	Symphony Foyer Entrance (Lobby Level)
8:00 am – 5:00 pm	Full-Day Short Courses	
8:00 am – noon	Morning Short Courses	
Noon – 1:00 pm	Lunch	
1:00 – 5:00 pm	Afternoon Short Courses	
6:00 – 9:00 pm	ICSA Board Meeting	Green Room (Lobby Level)

Online agenda: symposium2024.icsa.org/detailed-agenda

Course catalog: symposium2024.icsa.org/program/course-catalog

Ideas for lunch: symposium2024.icsa.org/nashville

Share your experience on social media! #ICSAnashville

C1: Deep Learning applications in statistical problems

8:00 AM–5:00 PM, Ocean Way (Mezzanine Level)

Instructors: Hongtu Zhu, University of North Carolina at Chapel Hill; Xiao Wang, Purdue University; Runpeng Dai, University of North Carolina at Chapel Hill

This short course delves into the intersection of Deep Learning and statistical analysis. Participants will explore and apply Deep Learning methodologies to tackle various statistical problems. The course covers advanced topics such as longitudinal data analysis, survival analysis, quantile regression, autoencoders, generative models, and handling spatial-temporal data using Deep Learning techniques.

C2: Statistical methods for time-to-event data from multiple sources: A causal inference perspective

8:00 AM–5:00 PM, Sound Emporium A/B (Mezzanine Level)

Instructors: Xiaofei Wang, Duke University School of Medicine; Shu Yang, North Carolina University

The short course will review important statistical methods for survival data arising from multiple data sources, including randomized clinical trials and observational studies. The entire short course consists of four parts and all parts will be discussed in a unified causal inference framework. In each part, we will review the theoretical background. Supplemented with data examples, the application of these methods in practice and implementation of these methods in freely available statistical software will be emphasized. Interactive sessions on implementing the new methods in R will be held. The methodology work related to the short course for both instructors has been funded by NIH R01 and FDA U01 grants.

C3: Structural equation modeling and its applications using R and SAS

8:00 AM–5:00 PM, RCA (Mezzanine Level)

Instructors: Din Chen, Arizona State University; Yiu-Fai Yung, SAS Institute

Originating from the social sciences, structural equation modeling (SEM) is becoming more popular in fields such as education, health science, and medical sciences. This one-day short course aims to provide an overview of SEM and to demonstrate its applications by using R and SAS software based on our newly published book, *Structural Equation Modeling Using R/SAS: A Step-by-Step Approach with Real Data Analysis* (Chapman and Hall/CRC, 2023). We will cover some main SEM topics, including path analysis, confirmatory factor analysis, mediation analysis in longitudinal settings, structural relations with latent variables, multiple-group SEM, latent growth-curve modeling, and model modification. Real-world applications are compiled to demonstrate its applications in social, educational, behavioral, and marketing research. Mathematical and statistical foundations of SEM are discussed at a level suitable for general understanding.

This course is designed for statisticians and data analysts who like to learn SEM techniques for their own research and applications. Both R package "lavaan" (latent variable analysis) and the CALIS procedure of SAS/STAT will be used to demonstrate model specifications, fitting, and result interpretations.

Attendees are expected to have a basic understanding of regression analysis. Experience using R and SAS software is not required for understanding the general SEM techniques.

C4: FUNCTIONAL DATA ANALYSIS AND ITS APPLICATIONS

8:00 AM–noon, Platinum (Lower Level)
Instructor: Pang Du, Virginia Tech

This course aims to introduce the modern field of functional data analysis to a general audience with the emphasis on how the relevant techniques can be applied to real examples. As a generalization of the traditional data concepts from numbers and vectors of numbers to curves and surfaces, functional data have attracted a lot of attention from statisticians and found many interesting applications in a variety of fields in the past decades. The course will start with the introduction of real examples for functional data. Based on these examples, common functional data analysis techniques such as function smoothing, functional principal component analysis, and functional linear regression models will be presented. R implementation of these techniques will be introduced and demonstrated.This course aims to introduce the modern field of functional data analysis to a general audience with the emphasis on how the relevant techniques can be applied to real examples. As a generalization of the traditional data concepts from numbers and vectors of numbers to curves and surfaces, functional data have attracted a lot of attention from statisticians and found many interesting applications in a variety of fields in the past decades. The course will start with the introduction of real examples for functional data. Based on these examples, common functional data analysis techniques such as function smoothing, functional principal component analysis, and functional linear regression models will be presented. R implementation of these techniques will be introduced and demonstrated.

C5: INFERENCE ON TREATMENT EFFECTS IN CLINICAL TRIALS WITH TERMINAL AND NON-TERMINAL EVENTS IN THE PRESENCE OF COMPETING RISKS

8:00 AM–noon, Melody (Lobby Level)
Instructor: Song Yang, Office of Biostatistics Research, NHLBI, NIH

Clinical trials often involve a terminal event (e.g., cardiovascular death) and some non-terminal events (e.g., stroke) where the terminal event may censor the non-terminal events and may also be subject to competing risks. The traditonal first event analysis does not use data fully and opaquely mixes different events, leading to wide CI and power loss. Various methods have been proposed in recent years, some involving complex models and others having intuitive appeal but hidden conditions. Furthermore, some are well-suited for etiological studies, while others are more convenient for testing and summarizing treatment effects. It is challenging to navigate the landscape in search of improved efficiency without biased, difficult to interpret, and non-generalizable results.

This user-friendly course addresses these challenges by discussing strengths and limitations of various approaches such as Copula models, multi-state models, restricted mean time, win ratio and their respective variants. Recommendations are given on which methods to use for a possible treatment effect scenario, emphasizing practical and robust analyses. Suggestions are made that facilitate the development of design and analysis plan for a future trial. Choices between hazard-oriented methods and cumulative incidence function-based methods and their alignment with clinical questions of interest are discussed. The methods are illustrated with Octave/MATLAB on data from a few recent large trials.

C6: UNLOCKING THE POWER OF SEMIPARAMETRIC MODELS: A PRACTICAL TUTORIAL FOR ANALYZING COMPLEX DATA WITH MINIMUM ASSUMPTIONS

8:00 AM–noon, Lyric (Lobby Level)

Instructors: Xin Tu, University of California, San Diego; Tuo Lin, University of Florida; Jinyuan Liu, Vanderbilt University Medical Center

This half-day short course will give biostatisticians and data scientists an engaging overview of semiparametric modeling via real-world applications with complex structures, such as high-throughput sequencing and network data. Both classical and cutting-edge semiparametric techniques will be explored, highlighting their roles in balancing robustness, flexibility, and efficiency with minimum assumptions.

The foundation of statistical inference relies on models with explicit or implicit assumptions about the underlying data-generating process. Often, these models are parametric, characterized by finite-dimensional parameters. They have only limited robustness in practice, which championed the advancement of semiparametric modeling that blends finite-dimensional parameters of interest with infinite-dimensional nuisance parameters. Such flexibility has led to emerging applications in many research disciplines, evidently focusing on causal inference, missing data, survival, and survey studies.

This short course will break into two halves. The first half introduces the fundamental concepts of semiparametric models and outlines their roles in robust inference without and with missing data. Some recent advances will be discussed in the second half, covering diverse applications that scale up to the high-dimensional microbiome data and HIV viral genetic linkage networks while also scaling down to inferences encountering outliers and small sample sizes.

C7: EVERYDAY REPRODUCIBILITY: SIMPLE FLEXIBLE TOOLS FOR MAKING ANALYSES MORE ACCESSIBLE AND REPRODUCIBLE

1:00 PM–5:00 PM, Gold (Lower Level)

Instructors: Gregory Hunt, William & Mary; Johann Gagnon-Bartsch, University of Michigan

Ensuring that analyses are reproducible is important for statisticians broadly from academics to industry. Indeed, the ability to reproduce third-party results is fundamental to the scientific process itself as well as to the public confidence in this process. In addition to ensuring that analyses are reproducible, it is important that these analyses can be easily shared, accessed, and explored. While critically important, building analyses that are computationally reproducible, shareable, and accessible is not a trivial task. This "reproducibility crisis" been recognized in popular science and professional statistical societies alike.

C8: INTRODUCTION TO NEUROIMAGE ANALYSIS FOR BIOSTATISTICIANS

1:00 PM–5:00 PM, Lyric (Lobby Level)

Instructors: Catie Chang, Vanderbilt University; Sarah Goodale, Vanderbilt University; Simon Vandekar, Vanderbilt University Medical Center

Beginning to work with neuroimaging data can be overwhelming for many biostatisticians whose methodological background could provide important tools and insights into modern challenges of neuroimage analysis. This course will provide a background of the data types, neuroimaging I/O and analysis resources, some preprocessed open source datasets, and recent concerns and challenges in the neuroimaging commu-

nity. The goal of the course is to provide biostatisticians with the resources to begin developing and implementing statistical methods that may be useful to the neuroimaging community. The course will be primarily didactic with a follow-along tutorial on loading and working with imaging data in R/Python/Matlab at the end.

C9: MODEL-ASSISTED DESIGNS: MAKE ADAPTIVE CLINICAL TRIALS EASY AND ACCESSIBLE

1:00 PM–5:00 PM, Platinum (Lower Level)

Instructors: Ying Yuan, University of Texas MD Anderson Cancer Center; Yong Zang, Indiana University

Drug development and clinical research face the challenges of prohibitively high costs, high failure rates, long trial duration, and slow accrual. One important approach to addressing this pressing issue is to use novel adaptive designs, which unfortunately can be hampered by the requirement of complicated statistical modeling, demanding computation, and expensive infrastructure for implementation.

This short course is designed to provide an overview of model-assisted designs, a new class of designs developed to simplify the implementation of adaptive designs in practice. Model-assisted designs are derived based on rigorous statistical theory, and thus possess superior operating characteristics and great flexibility, while can be implemented as simply as algorithm-based designs. Easy-to-use Shiny applications on the web and downloadable standalone programs will be introduced to facilitate the study design and conduct. The main application areas include adaptive dose-finding, adaptive toxicity and efficacy evaluation, posterior probability and predictive probability for interim monitoring of study endpoints, outcome-adaptive randomization, hierarchical models, multi-arm, multi-stage designs, and platform designs. Lessons learned from real trial examples and practical considerations for conducting adaptive designs will be given. couses

C10: INTRODUCTION OF BIOMARKER DISCOVERY IN CANCER RESEARCH

1:00 PM–5:00 PM, Melody (Lobby Level)

Instructors: Xiaoli Zhang, The Ohio State University; Lianbo Yu, The Ohio State University

With the significant advancements in genomic profiling technologies and the emergence of selective molecular targeted therapies, biomarkers have played an increasingly pivotal role in both the prognosis and treatment of various diseases, most notably cancer. This workshop is designed to begin with an introductory overview of basic concepts of biomarkers, the diverse categories of biomarkers, and commonly employed biotechnologies for biomarker detection, with a special focus on gene mutations and gene expression. Furthermore, we will discuss processes of biomarker discovery and development, outlining the key steps involved and the current analytical methodologies utilized. Following this, we will discuss the identification of driver gene mutations and altered gene expression, using lung cancer data in The Cancer Genome Atlas (TCGA) as an illustrative example with using SAS or R code as practical demonstrations to enhance understanding. In the latter part of this workshop, we will discuss commonly utilized biostatistics and bioinformatics tools, including data visualization, survival analysis and classification methods, which are employed to predict disease progression and patient survival outcomes based on these critical biomarkers. If time permits, we will also discuss the concept and analysis of scRNA-seq data. By the conclusion of this course, participants will have acquired a broad and fundamental understanding of biomarker discovery in cancer research.

Monday, June 17, 2024

7:30 am – 6:30 pm	Registration	Symphony Foyer Entrance (Lobby Level)
7:00 am – 8:00 am	Breakfast	Symphony Foyer (Lobby Level)
8:00 – 8:30 am	Welcome and Opening Remarks	Symphony 2&3 (Lobby Level)
8:30 – 9:30 am	Keynote: Hongtu Zhu, PhD, FASA, FIMS	Symphony 2&3 (Lobby Level)
9:30 – 10:00 am	Coffee Break	Symphony Foyer (Lobby Level)
10:00 – 11:30 am	Invited Sessions	
11:30 am – 1:00 pm	Lunch	
1:00 – 2:30 pm	Invited Sessions and Panels	
2:30 – 3:00 pm	Coffee Break	Symphony Foyer (Lobby Level)
3:00 – 4:30 pm	Invited Sessions and Panels	
5:00 – 6:00 pm	ICSA General Member Meeting	Symphony 1 (Lobby Level)
6:00 – 8:00 pm	Poster Session and Mixer	Starstruck Gallery (Mezzanine Level)

Online agenda: symposium2024.icsa.org/detailed-agenda

Ideas for lunch: symposium2024.icsa.org/nashville

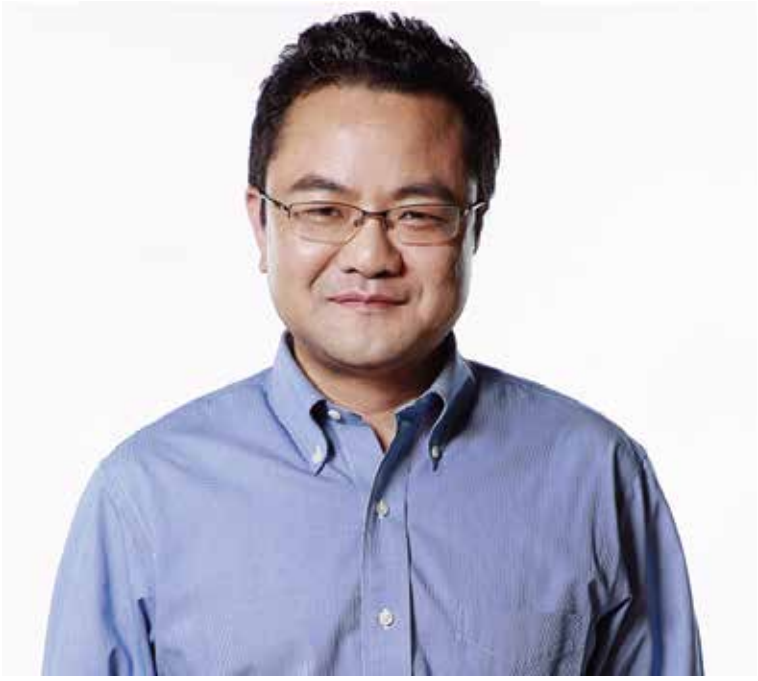
Share your experience on social media! #ICSAnashville

Keynote Talk - Monday, June 17

Uniting statistics and AI for revolutionizing medical data analysis and more

Hongtu Zhu, PhD, FASA, FIMS

8:30 am, Symphony 2&3 (Lobby Level)



This talk provides an insightful overview of integrating artificial intelligence (AI) and statistical methods in medical data analysis. It is structured into three key sections:

Introduction to Medical Image Data Analysis: This section sets the stage by outlining the fundamentals and significance of medical image analysis in healthcare, charting its evolution and current applications.

State-of-the-Art AI Applications and Statistical Challenges: Here, we explore the impact of AI, particularly deep learning, on medical imaging, and address the accompanying statistical challenges, such as data quality and model interpretability.

Opportunities for Statisticians: The final section highlights the critical role of statisticians in refining AI applications in medical imaging, focusing on opportunities for advancing algorithmic accuracy and integrating statistical rigor.

The talk aims to demonstrate the crucial synergy between AI and statistics in enhancing medical data analysis, emphasizing the evolving challenges and the vital contributions of statisticians in this domain.

Dr. Hongtu Zhu is professor of biostatistics, statistics, radiology, computer science, and genetics at the University of North Carolina at Chapel Hill. He was a DiDi Fellow and Chief Scientist of Statistics at DiDi Chuxing between 2018 and 2020 and held the Endowed Bao-Shan Jing Professorship in Diagnostic Imaging at MD Anderson Cancer Center between 2016 and 2018. He is an internationally recognized expert in statistical learning, medical image analysis, precision medicine, biostatistics, artificial intelligence, and big data analytics. He received an established investigator award from the Cancer Prevention Research Institute of Texas in 2016 and received the INFORMS Daniel H. Wagner Prize for Excellence in Operations Research Practice in 2019. He has published more than 300 papers in top journals, including *Nature*, *Science*, *Cell*, *Nature Genetics*, *Proceedings of the National Academy of Sciences* (PNAS), *Annals of Statistics* (AOS), *Journal of the American Statistical Association* (JASA), and *Journal of the Royal Statistical Society* (JRSSB), as well as presenting 50+ conference papers at top conferences, including meetings for Neural Information Processing Systems (NeurIPS), Association for the Advancement of Artificial Intelligence (AAAI), Knowledge Discovery and Data Mining (KDD), International Conference on Data Mining (ICDM), Medical Image Computing and Computer Assisted Intervention (MICCAI) Society, and Intelligent Platform Management Interface (IPMI).

S1: STATISTICAL METHODS FOR SINGLE CELL AND SPATIAL OMICS DATA

Monday, June 17, 2024
10:00 AM–11:30 AM, Symphony 1 (Lobby Level)

Organizer: Hongkai Ji, Johns Hopkins University
Chair: Hongkai Ji, Johns Hopkins University

10:00 AM–10:20 AM Speaker: Guo-Cheng Yuan, Icahn School of Medicine at Mount Sinai

Characterizing spatially continuous variations in tissue microenvironment through niche trajectory analysis

Author(s): Wen Wang, Icahn School of Medicine at Mount Sinai; Shiwei Zheng, Icahn School of Medicine at Mount Sinai; Sujung Shin, Icahn School of Medicine at Mount Sinai; Guo-Cheng Yuan, Icahn School of Medicine at Mount Sinai

Recent technological developments have made it possible to map the spatial organization of a tissue at the single-cell resolution. However, computational methods for analyzing spatially continuous variations in tissue microenvironment are still lacking. Here we present ONTraC as a strategy that constructs niche trajectories using a graph neural network-based modeling framework. Our benchmark analysis shows that ONTraC performs more favorably than existing methods for reconstructing spatial trajectories. Applications of ONTraC to public spatial transcriptomics datasets successfully recapitulated the underlying anatomical structure, and further enabled detection of tissue microenvironment-dependent changes in gene regulatory networks and cell-cell interaction activities during embryonic development. Taken together, ONTraC provides a useful and generally applicable tool for the systematic characterization of the structural and functional organization of tissue microenvironments.

10:20 AM–10:20 AM Speaker: Wei Vivian Li, University of California, Riverside

spVC for the detection and interpretation of spatial gene expression variation

Author(s): Shan Yu, University of Virginia; Wei Vivian Li, University of California, Riverside

Recent advances in spatially resolved transcriptomics technologies have opened up new avenues for understanding gene expression heterogeneity in spatial contexts, and an important task in spatial transcriptomics analysis is the identification of spatially variable genes (SVGs). While various computational methods exist for SVG detection, most focus solely on statistical significance and have limitations in capturing continuous expression patterns across spatial domains and incorporating cell/spot-level covariates.

To address these challenges, we introduce spVC, a novel statistical method to detect and interpret SVGs based on a generalized Poisson model. spVC integrates constant and spatially varying effects of cell/spot-level covariates, enabling comprehensive exploration of gene expression variability and enhancing interpretability. It provides a convenient tool to identify potential factors that contribute to gene expression variability, including spatial locations and other cell/spot-level covariates such as cell types or tissue layers. It offers estimation and statistical inference tools for both constant and spatially varying coefficients, allowing for the selection of different types of SVGs. In summary, spVC is a versatile tool for the identification, interpretation, and comprehension of gene expression variation in spatial transcriptomics data.

Dynamic prediction models capable of retaining accuracy by evolving over time could play a significant role for monitoring disease progression in clinical practice. In biomedical studies with long-term follow up, participants are often monitored through periodic clinical visits with repeat measurements until an occurrence of the event of interest (e.g., disease onset) or the study end. Acknowledging the dynamic nature of disease risk and clinical information contained in the longitudinal markers, we propose an innovative concordance-assisted learning algorithm to derive a real-time risk stratification score. The proposed approach bypasses the need to fit regression models, such as joint models of the longitudinal markers and time-to-event outcome, and hence enjoys the desirable property of model robustness. Simulation studies confirmed that the proposed method has satisfactory performance in dynamically monitoring the risk of developing disease and differentiating high-risk and low-risk population over time. We apply the proposed method to the Alzheimer's Disease Neuroimaging Initiative data and develop a dynamic risk score of Alzheimer's Disease for patients with mild cognitive impairment using multiple longitudinal markers and baseline prognostic factors.

10:40 AM–11:00 AM Speaker: Kevin Lin, University of Washington

eSVD-DE: Cohort-wide differential expression in single-cell RNA-seq data using exponential-family embeddings

Author(s): Kevin Lin, University of Washington

Single-cell RNA-sequencing (scRNA) datasets are becoming increasingly popular in clinical and cohort studies. Still, there is a lack of methods to investigate differentially expressed (DE) genes among such datasets with numerous individuals. While multiple methods exist to find DE genes for scRNA data from limited individuals, differential-expression testing for large cohorts of case and control individuals using scRNA data poses unique challenges due to substantial effects of human variation—i.e., individual-level confounding covariates that are difficult to account for in the presence of sparsely-observed genes. In this talk, I develop the eSVD-DE, a matrix factorization that pools information across genes and removes confounding covariate effects, followed by a novel two-sample test in mean expression between case and control individuals. In general, differential testing after dimension reduction yields an inflation of Type-1 errors. However, we overcome this by testing for differences between the case and control individuals' posterior mean distributions via a hierarchical model. In previously published datasets of various biological systems, eSVD-DE has more accuracy and power than other DE methods typically repurposed for analyzing cohort-wide differential expression. Altogether, accurately identifying differential expression on the individual level, instead of the cell level, is important for linking scRNA-seq studies to our understanding of the human population.

11:00 AM–11:20 AM Speaker: Sunduz Keles, University of Wisconsin-Madison

Learning genomic multi-way interactions from single-cell chromatin conformation capture data
Author(s): Sunduz Keles, University of Wisconsin, Madison; Kwangmoon Park, University of Wisconsin, Madison

A number of foundational analysis methods have emerged for single cell chromatin conformation (scHi-C) datasets capturing 3D organizations of genomes at the single cell level; however, these scHi-C datasets are currently under-utilized. We introduce ELECT to infer multi-way interactions (e.g., looping among multiple genomic elements such as multiple enhancers of a gene) from scHi-C data. ELECT employs a Dirichlet-multinomial spline model, incorporates well-known genomic distance bias of the Hi-C data, and facilitates unbiased inference for multi-way interactions by aggregating locus-pair level summary statistics. We applied ELECT to both low and high resolution scHi-C datasets and carried out evaluations with external genomic and epigenomic resources including data from SPRITE, scNanoHi-C, multiplexed DNA-FISH, and DNA methylation assays. Application of ELECT to human brain scHi-C revealed multi-way interactions that involved GWAS SNPs associated with psychiatric disorders, including autism and schizophrenia.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S2: APPLICATION, THEORY, AND COMPUTING OF STATISTICAL LEARNING

Monday, June 17, 2024
10:00 AM–11:30 AM, Blackbird A (Mezzanine Level)

Organizer: Yu Cheng, University of Pittsburgh
Chair: Jing Ning, University of Texas MD Anderson Cancer Center

10:00 AM–10:20 AM Speaker: Grace Yi, University of Western Ontario

Correcting misclassification effects on Q-learning for dynamic treatment regimes

Author(s): Grace Yi, University of Western Ontario

Research on dynamic treatment regimes has enticed extensive interest. Many methods have been proposed in the literature, which, however, are vulnerable to the presence of misclassification in covariates. In particular, although Q-learning has received considerable attention, its applicability to data with misclassified covariates is unclear. In this article, we investigate how ignoring misclassification in binary covariates can impact the determination of optimal decision rules in randomized treatment settings, and demonstrate its deleterious effects on Q-learning through empirical studies. We present two correction methods to address misclassification effects on Q-learning. Numerical studies reveal that misclassification in covariates induces non-negligible estimation bias and that the correction methods successfully ameliorate bias in parameter estimation.

10:20 AM–10:20 AM Speaker: Wen Li, University of Texas Health Science Center

Dynamic and concordance-assisted learning for risk stratification with application to Alzheimer's disease

Author(s): Wen Li, University of Texas Health Science Center; Ruosha Li, University of Texas Health Science Center; Ziding Feng, Fred Hutchinson Cancer Research Center; Jing Ning, University of Texas MD Anderson Cancer Center

Dynamic prediction models capable of retaining accuracy by evolving over time could play a significant role for monitoring disease progression in clinical practice. In biomedical studies with long-term follow up, participants are often monitored through periodic clinical visits with repeat measurements until an occurrence of the event of interest (e.g., disease onset) or the study end. Acknowledging the dynamic nature of disease risk and clinical information contained in the longitudinal markers, we propose an innovative concordance-assisted learning algorithm to derive a real-time risk stratification score. The proposed approach bypasses the need to fit regression models, such as joint models of the longitudinal markers and time-to-event outcome, and hence enjoys the desirable property of model robustness. Simulation studies confirmed that the proposed method has satisfactory performance in dynamically monitoring the risk of developing disease and differentiating high-risk and low-risk population over time. We apply the proposed method to the Alzheimer's Disease Neuroimaging Initiative data and develop a dynamic risk score of Alzheimer's Disease for patients with mild cognitive impairment using multiple longitudinal markers and baseline prognostic factors.

10:40 AM–11:00 AM Speaker: Linxi Liu, University of Pittsburgh

Posterior concentration rates for unsupervised trees and forests

Author(s): Linxi Liu, University of Pittsburgh; Li Ma, Duke University

Tree-based methods are popular nonparametric tools for capturing spatial heterogeneity and making predictions in multivariate problems. In unsupervised learning, trees and their ensembles have also been applied to a wide range of statistical inference tasks, such as multi-resolution sketching of distributional variations, localization of high-density regions, and design of efficient data compression schemes. In this talk, we will focus on the density estimation problem—a fundamental one in unsupervised learning. We consider the optional Pólya tree (Wong and Ma, 2010) prior or its variations on individual trees. First we show that Bayesian density trees can achieve minimax (up to a logarithmic term) convergence over the anisotropic Besov class, which implies that tree based methods can adapt to spatially inhomogeneous features of the underlying density function, and can achieve fast convergence as the dimension increases. We will also introduce a novel Bayesian model for forests, and show that for a class of Hölder continuous functions, such type of density forests can achieve faster convergence than trees. The convergence rate is adaptive in the sense that to achieve such a rate we do not need any prior knowledge of the smoothness level of the density. The Bayesian framework naturally provides a stochastic search algorithm over either the tree space or the forest one. For both Bayesian density trees and forests, we will provide several numerical results to illustrate their performance in the moderately high-dimensional case.

11:00 AM–11:20 AM Speaker: Jun Yan, University of Connecticut

Optimal subsampling for semi-parametric accelerated failure time models with massive survival data using a rank-based approach

Author(s): Zehan Yang, University of Connecticut; HaiYing Wang, University of Connecticut; Jun Yan, University of Connecticut

Subsampling is a practical strategy for analyzing vast survival data, which are progressively encountered across diverse research domains. While the optimal sub-sampling method has been applied to inferences for Cox models and parametric accelerated failure time (AFT) models, its application to semi-parametric AFT models with rank-based estimation has received limited attention. The challenges arise from the non-smooth estimating function for regression coefficients and the seemingly zero contribution from censored observations in estimating functions in the commonly seen form. To address these challenges, we develop optimal subsampling probabilities for both event and censored observations by expressing the estimating functions through a well-defined stochastic process. Meanwhile, we apply an induced smoothing procedure to the non-smooth estimating functions. As the optimal sub-sampling probabilities depend on the unknown regression coefficients, we employ a two-step method to obtain a feasible estimation procedure. An additional benefit of the method is its ability to resolve the issue of underestimation of the variance when the subsample size approaches the full sample size. We validate the performance of our estimators through a simulation study and apply the methods to analyze the survival time of lymphoma patients in the Surveillance, Epidemiology, and End Results program.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S3: STATISTICAL LEARNING AND COMPUTATIONAL TOOLS FOR BIG BIOMEDICAL DATA

Monday, June 17, 2024
10:00 AM–11:30 AM, Lyric (Lobby Level)

Organizer: Bingxin Zhao, University of Pennsylvania
Chair: Xiaochen Yang, Purdue University

10:00 AM–10:20 AM Speaker: Jin Jin, University of Pennsylvania

Mendelian randomization analysis of latent exposures co-regulating multiple correlated traits

Author(s): Yue Yu, University of Pennsylvania; Jin Jin, University of Pennsylvania

Mendelian Randomization (MR) is an increasingly popular tool for conducting causal inference using observational data from genome-wide association studies (GWAS). In some settings, we want to identify causal signal among a set of correlated traits, such as phenotypes associated with different subtypes of a disease, in other words, the underlying exposure of interest (e.g., systematic inflammation) may not be directly observable, but measurements can be available on multiple traits that are coregulated by the exposure. We propose methods for conducting MR analysis on latent exposures, which test the significance for, and the direction of, the effect of a latent exposure(s) by leveraging information from multiple related traits. Simulation studies show that the proposed methods have well-controlled type I error rates and enhanced power compared to singletrait MR tests under various types of pleiotropy. Applications based on genetic association statistics across five inflammatory biomarkers provide evidence for potential causal effects of inflammation on increasing the risk of coronary artery disease, colorectal cancer, and rheumatoid arthritis, while standard MR analysis for individual biomarkers fails to detect consistent evidence for such effects.

10:20 AM–10:20 AM Speaker: Bingxin Zhao, University of Pennsylvania

Exploring cross-trait genetic architectures: The BIGA platform

Author(s): Bingxin Zhao, University of Pennsylvania

Summary statistics derived from extensive genetic and genomic studies offer rich insights for exploring shared genetic architectures among phenotypes across different studies and cohorts. Yet, the systematic analysis of these massive summary statistics poses significant logistical and computational challenges. In this talk, I will introduce the BIGA platform (bigagwas.org), a website providing unified data analysis pipelines and centralized data resources. Our team has created a framework that applies statistical genetics tools in a cloud computing environment, integrating it with extensive curated genome-wide association studies (GWAS) datasets. Through BIGA, users can upload data, submit jobs, and share results, providing the research community with a convenient tool for consolidating GWAS data and generating new insights. This is a joint work with Fei Xue and Yujue Li

10:40 AM–11:00 AM Speaker: Hai Shu, New York University

DeepFDR: A Deep Learning-based False Discovery Rate control method for neuroimaging data
Author(s): Hai Shu, New York University; Taehyo Kim, New York University; Qiran Jia, University of Southern California; Mony de Leon, Weill Cornell Medicine

Voxel-based multiple testing is widely used in neuroimaging data analysis. Traditional false discovery rate (FDR) control methods often ignore the spatial dependence among the voxel-based tests and thus suffer from substantial loss of testing power. While recent spatial FDR control methods have emerged, their validity and optimality remain questionable when handling the complex spatial dependencies of the brain. Concurrently, deep learning methods have revolutionized image segmentation, a task closely related to voxel-based multiple testing. In this paper, we propose DeepFDR, a novel spatial FDR control method that leverages unsupervised deep learning-based image segmentation to address the voxel-based multiple testing problem. Numerical studies, including comprehensive simulations and Alzheimer's disease FOG-PET image analysis, demonstrate DeepFDR's superiority over existing methods. DeepFDR not only excels in FDR control and effectively diminishes the false nondiscovery rate, but also boasts exceptional computational efficiency highly suited for tackling large-scale neuroimaging data.

11:00 AM–11:20 AM Speaker: Chong Wu, University of Texas MD Anderson Cancer Center

PURE: An integrated approach for causal protein discovery leveraging cis- and trans-acting elements
Author(s): Zichen Zhang, University of Texas MD Anderson Cancer Center; Lang Wu, University of Hawaii; Bingxin Zhao, University of Pennsylvania; Chong Wu, University of Texas MD Anderson Cancer Center

An enhanced comprehension of genetic regulation of the proteome can accelerate the elucidation of causal mechanisms for complex traits. Proteome-wide association studies (PWAS), which leverage protein quantitative trait loci (pQTL) datasets integrated with genome-wide association studies (GWAS), are frequently used to identify probable causal proteins. Analogous to transcriptome-wide association studies (TWAS), current PWAS methods predominantly focus on cis-acting elements, constructing protein levels prediction models using individual-level pQTL datasets, often limited by sample size. However, trans-acting elements can account for a significant proportion of variation in many protein markers and often play essential regulatory roles. To maximize the potential of transacting elements and summary-level pQTL data for improving the robustness and power of PWAS, we introduce a novel PWAS method, referred to as Protein-trait association Using cis- and tRansregulation Estimation (PURE). Our method involves two steps. First, we develop prediction models using each cis- and trans-acting element leveraging summary-level data with large sample size, thereby addressing the relatively weak effects of trans-acting elements. Next, we estimate the associations between phenotype and genetically predicted protein levels for each cis- and trans-acting element. These associations, estimated from each locus, are then combined using an iterative algorithm to account for certain outliers and randomness in outlier detection. Applying our novel method to deCODE summary-level data, encompassing plasma protein levels of 4,907 proteins derived from 35,559 Icelanders, we constructed 2,127 protein prediction models demonstrating satisfactory performance (R2>0.01), achieving a 58% improvement over the existing model constructed using the ARIC dataset. Further external validation of our models using the INTERVAL data yielded a high validation rate. Finally, in a case study for Alzheimer's disease using recent GWAS data

of 111,326 clinically diagnosed/proxy cases and 677,663 controls, our model identified 207 likely causal proteins under Bonferroni correction. In contrast, competing methods identified 19 likely causal proteins. We will release a companion software on GitHub to enable wider use of our method.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S4: HIGHLIGHTS OF A SPECIAL ISSUE FOR JOURNAL OF BIOPHARMACEUTICAL STATISTICS: ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING USE CASES IN PHARMACEUTICAL INDUSTRY

Monday, June 17, 2024
10:00 AM–11:30 AM, Ocean Way (Mezzanine Level)

Organizer: Margaret Gamalo, Pfizer, Inc.
Chair: Yushi Liu, Eli Lilly and Company

10:00 AM–10:20 AM Speaker: Dongyan Yan, Eli Lilly and Company

scRAA: The development of a robust and automatic annotation procedure for single-cell RNA sequencing data

Author(s): Dongyan Yan, Eli Lilly and Company; Zhe Sun, Eli Lilly and Company; Jiyuan Fang, Eli Lilly and Company; Shanshan Cao, Eli Lilly and Company; Wenjie Wang, Eli Lilly and Company; Xinyue Chang, Eli Lilly and Company; Sarkhan Badirli, Eli Lilly and Company; Haoda Fu, Eli Lilly and Company; Yushi Liu, Eli Lilly and Company

A critical task in single-cell RNA sequencing (scRNA-seq) data analysis is to identify cell types from heterogeneous tissues. While the majority of classification methods demonstrated high performance in scRNA-seq annotation problems, a robust and accurate solution is desired to generate reliable outcomes for downstream analyses, for instance, marker genes identification, differentially expressed genes, and pathway analysis. It is hard to establish a universally good metric. Thus, a universally good classification method for all kinds of scenarios does not exist. In addition, reference and query data in cell classification are usually from different experimental batches, and failure to consider batch effects may result in misleading conclusions. To overcome this bottleneck, we propose a robust ensemble approach to classify cells and utilize a batch correction method between reference and query data. We simulated four scenarios that comprise simple to complex batch effect and account for varying cell type proportions. We further tested our approach on both lung and pancreas data. We found improved prediction accuracy and robust performance across simulation scenarios and real data. The incorporation of batch effect correction between reference and query, and the ensemble approach improve cell type prediction accuracy while maintaining robustness. We demonstrated these through simulated and real scRNA-seq data.

10:20 AM–10:20 AM Speaker: Rachael Liu, Takeda Pharmaceuticals

DOD-BART: Machine learning-based Dose Optimization Design incorporating patient-level prognostic factors via Bayesian Additive Regression Trees

Author(s): Rachael Liu, Takeda; Yunqi Zhao, Takeda; Jianchang Lin, Takeda; Andy Chi, Takeda; Simon Davies, Takeda

Dose optimization is a critical stage of drug development in oncology and other disease areas. Early phase clinical trials are inherently heterogeneous due to their exploratory nature. The process of identifying an optimal dose involves careful considerations of the patient population, evaluation of therapeutic potential, and exploration of the dose-response and dose-toxicity relationships to ensure that it is safe and effective for the intended use. However, the complex

mechanism of actions and uncertainties during dose optimization often introduce substantial gaps between those early phase trials and phase 3 randomized control trials. These gaps can indeed increase the chances of failure. To address these challenges, we propose a novel seamless phase I/II design, namely DOD-BART design, which utilizes machine learning technique, specifically Bayesian Additive Regression Trees (BART) to fully incorporate patient-level prognostic factors and outcomes. Our design provides a streamlined approach for dose exploration and optimization, automatically updated with emerging data to allocate patients to the most promising dose levels. DOD-BART elucidates disease relationships, analyzes and synthesizes emerging data, augments operational efficiency, and guides dose optimization for suitable population. Simulation studies demonstrate the robust performances of the DOD-BART designs across a variety of realistic settings, with high probabilities of correctly identifying the optimal dose, allocating patients more to tolerable and efficacious dose levels, making less biased estimates, and efficiently utilizing patients' data.

10:40 AM–11:00 AM Speaker: Yu Deng, AbbVie Inc.

Developing large language models for adverse drug event detection in tweets

Author(s): Yu Deng, AbbVie Inc.; Yunzhao Xing, AbbVie Inc.; Sheng Zhong, AbbVie Inc.; Li Wang, AbbVie Inc.

Adverse drug events (ADEs) are one of the major causes of hospital admissions and are associated with increased morbidity and mortality. Post-marketing ADE identification is one of the most important phases of drug safety surveillance. Traditionally, data sources for post-marketing surveillance mainly come from spontaneous reporting system such as the Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS). Social media data such as tweets contains rich patient and medication information and could potentially accelerate drug surveillance research. However, ADE information is usually locked in the text in social media data, making it difficult to be employed by traditional statistical approaches. Natural language processing methods can be used to mine valuable information from texts. In recent years, transformer-based large language models (LLMs) have shown promise in many medical-related tasks. In this study, we developed several LLMs to perform ADE classification in Twitter data. We fine-tuned various BERT-based models, including BERT-base, Bio ClinicalBERT, RoBERTa, and RoBERTa-large. We also experimented with generative models, including ChatGPT few-shot prompting and ChatGPT fine-tuned on the whole training data. Random search was implemented to find the optimal parameter settings for each LLM. We then evaluated the model performance based on sensitivity, specificity, negative predictive value, positive predictive value, accuracy, F1-measure, and area under the ROC curve (AUC). Our results showed that RoBERTa-large achieved the best F1-measure (0.73) among all the BERT based models. The ChatGPT fine-tuned model outperformed all models (F1 =0.75). Our feature importance analysis, based on 500 random samples and RoBERTa-Large, showed the most important features are: "headaches," "through," "restless," "nightmares," and "withdrawal." The good model performance and clinically relevant features show the potential of LLMs in augmenting ADE detection for post-marketing drug safety surveillance.

11:00 AM–11:30 AM **Q&A and Floor Discussion**

S5: NEW METHODS FOR WEARABLE DATA IN LONGITUDINAL STUDIES

Monday, June 17, 2024
10:00 AM–11:30 AM, Platinum (Lower Level)

Organizer: Luo Xiao, North Carolina State University
Chair: Luo Xiao, North Carolina State University

10:00 AM–10:20 AM Speaker: Erjia Cui, University of Minnesota

Fast univariate inference for large-scale physical activity studies

Author(s): Erjia Cui, University of Minnesota

In the realm of functional mixed models, fast marginal inferential techniques have emerged to model longitudinal Gaussian and non-Gaussian functional data. Among the multitude of methods available, Fast Univariate Inference (FUI) stands out for its considerable computational advantages when dealing with large-scale, high-dimensional datasets commonly encountered in physical activity studies. Recognizing the intricate nature of these data, we expand upon the original FUI approach proposed by Cui et al. (2022) to accommodate more complex data structures. The enhanced FUI methods are applied to several physical activity studies to uncover the temporal and longitudinal impact of key variables on physical activity. Methods are accompanied by R software.

10:20 AM–10:20 AM Speaker: Chongzhi Di, Fred Hutchinson Cancer Center

Robust multilevel functional principal component analysis with application to accelerometry data
Author(s): Chongzhi Di, Fred Hutchinson Cancer Center; Ken Wang, US Food and Drug Administration

In this talk, we discuss a new robust functional principal component analysis (FPCA) approach designed for multilevel functional data. Standard FPCA approaches might not perform well in the presence of heavy-tail distributions and outliers. We propose a more robust procedure based on pairwise spatial signs. We will discuss theoretical properties of the proposed procedure and demonstrate the empirical performance through extensive simulations. The proposed method will be illustrated by an application to accelerometry data in a large-scale epidemiological study.

10:40 AM–11:00 AM Speaker: Kristin Linn, University of Pennsylvania

Estimation and evaluation of individualized treatment rules following multiple imputation
Author(s): Jenny Shen, University of Pennsylvania; Rebecca Hubbard, University of Pennsylvania; Kristin Linn, University of Pennsylvania

Data-driven optimal treatment strategies promise to benefit patients, care providers, and other stakeholders by improving clinical outcomes and lowering healthcare costs. A treatment decision rule is a function that inputs patient-level information and outputs a recommended treatment. An important focus of precision medicine is to develop optimal treatment decision rules that maximize a population level distributional summary such as the expected value of a clinical outcome. However, guidance for estimating and evaluating optimal treatment decision rules in the presence of missing data is fairly limited. Our work is motivated by the Social Incentives to Encourage Physical Activity and Understand Predictors (STEP UP) study. In this trial, participants were randomized to a control arm or one of multiple interventions that were designed

to increase physical activity. Study participants were given wearable devices which were used to record daily step counts as a measure of physical activity. Many participants were missing at least one daily step count during the study period, and the missingness pattern within individuals was often non-monotone. We propose two frameworks for estimation and evaluation of an optimal treatment decision rule following multiple imputation and compare performance of the frameworks using simulated data. We apply our methods to the STEP UP data to determine whether a personalized intervention strategy might be expected to increase physical activity. more than the intervention that had the largest estimated average treatment effect

11:00 AM–11:20 AM Speaker: Xinkai Zhou, Johns Hopkins University

Analysis of Active/Inactive Patterns in the NHANES Data using generalized multilevel functional principal component analysis

Author(s): Xinkai Zhou, Johns Hopkins University; Julia Wrobel, Emory University; Ciprian Crainiceanu, Johns Hopkins University; Andrew Leroux, Colorado School of Public Health

Between 2011 and 2014 NHANES collected objectively measured physical activity data using wrist-worn accelerometers for tens of thousands of individuals for up to seven days. Here we analyze the minute-level indicators of being active, which can be viewed as binary (because there is an active indicator at every minute), multilevel (because there are multiple days of data for each study participant), functional (because within-day data can be viewed as a function of time) data. To extract within- and between-participant directions of variation in the data, we introduce Generalized Multilevel Functional Principal Component Analysis (GM-FPCA), an approach based on the dimension reduction of the linear predictor. Scores associated with specific patterns of activity are shown to be strongly associated with time to death. In particular, we confirm that increased activity is associated with time to death, a result that has been reported on other data sets. In addition, our method shows the previously unreported finding that maintaining a consistent day-to-day routine is strongly associated with a reduced risk of mortality (p-value < 0.001) even after adjusting for traditional risk factors. Extensive simulation studies indicate that GM-FPCA provides accurate estimation of model parameters, is computationally stable, and is scalable in the number of study participants, visits, and observations within visits. R code for implementing the method is provided.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S6: ADVANCEMENTS AND INNOVATIVE APPLICATIONS OF DATA/INFORMATION INTEGRATION METHODS IN BIOMEDICAL RESEARCH

Monday, June 17, 2024
10:00 AM–11:30 AM, Sound Emporium A/B (Mezzanine Level)

Organizer: Chixiang Chen, University of Maryland, School of Medicine
Chair: Chixiang Chen, University of Maryland School of Medicine

10:00 AM–10:20 AM Speaker: Yiwang Zhou, St. Jude Children's Research Hospital

Longitudinal self-learning of individualized treatment rules in a nutrient supplementation trial with missing data

Author(s): Yiwang Zhou, St. Jude Children's Research Hospital; Peter Song, University of Michigan

Longitudinal outcomes are pervasive in clinical studies where missing data may make derivation of individualized treatment rules (ITR) a more challenging task. We analyze a longitudinal calcium supplementation trial and establish a novel ITR to reduce the risk of lead exposure on children's development. Exposure to lead can seriously harm children's cognitive and neurobehavioral development, which necessitates clinical interventions, such as calcium supplementation intake during pregnancy. Using longitudinal outcomes from a calcium supplement trial, we develop a new ITR of daily calcium intake during pregnancy to mitigate persistent lead exposure by children of 3 years old. To overcome technical challenges arising from missing data, we propose a new learning method, termed longitudinal self-learning (LS-learning), that utilizes longitudinal measurements of a child's blood lead concentration in ITR derivation. Our LS-learning method relies on a weighted self-learning paradigm to synergize correlated training data sources. The resulting ITR helps lower expected blood lead concentration in children aged 0-3 years old should it be implemented in the whole study population.

10:20 AM–10:20 AM Speaker: Travis Canida, University of Maryland

High-dimension-to-high-dimension Bayesian variable selection for the fine mapping of phenome-wide transcriptome-wide association studies

Author(s): Travis Canida, University of Maryland; Hongjie Ke, University of Maryland; Zhenyao Ye, University of Maryland; Tianzhou Ma, University of Maryland

Transcriptome-wide association studies use predicted expression to identify genes associated with complex traits. As the number of traits increases nowadays, it becomes common to also perform phenome-wide association studies to identify associations between genes and a wide range of closely related phenotypes. Further, we may want to fine-map the associated genomic regions to identify the most causal genes that point to molecular mechanisms behind the associations. However, the high-dimensionality and complex correlation in both phenome and transcriptome data hinder the direct application of any existing variable selection methods to this problem. In this paper, we propose a novel high-dimension-to-high-dimension Bayesian variable selection method for the fine mapping of phenome-wide transcriptome-wide association studies. Simulations showed our method is advantageous in recovering the phenotype factor loadings and identifying the true set of causal genes, especially when both phenotypes and genes are of high dimension. We applied our method to UK Biobank and identified critical

causal genes for various phenotypes related to cognitive and physical functions.

10:40 AM–11:00 AM Speaker: Bryan Shepherd, Vanderbilt University

Synthetic HIV cohort data to bridge open science and privacy protections

Author(s): Zhuohui Liang, Vanderbilt University; Chao Yan, Vanderbilt University; Yanink Caro-Vega, Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran; Peter Rebeiro, Vanderbilt University; Stephany Duda, Vanderbilt University; Bradley Malin, Vanderbilt University; Bryan Shepherd, Vanderbilt University

Open research often clashes with data privacy laws and regulations, especially with respect to international health data. Synthetic data offers a viable middle-ground solution to enable sharing data that resemble the original data while mitigating privacy concerns. Although synthetic data can never fully replace the original data, we expect that synthetic data will be invaluable for training, making research quasi-reproducible, improving research efficiency, and generating hypotheses. In this study, we created a synthetic dataset for the Caribbean, Central and South America network for HIV epidemiology (CCASAnet) using generative adversarial network (GAN) techniques. A GAN is an unsupervised machine learning method that optimizes synthetic data generation through a real vs. synthetic data discrimination process. We used the CCASAnet cohort (59,208 people with HIV (PWH), enrolled 1991-2022) to simulate synthetic 'wide' (i.e., one row per PWH) cross-sectional data that contain 13 dates (e.g., dates of antiretroviral therapy (ART) initiation and death) and 37 landmark values (e.g., birth sex, CD4 at cohort entry), but do not contain longitudinal measurements (e.g., time-updated CD4). The synthetic data effectively captured the real data's marginal distributions. For example, the correlation between the synthetic and real data prevalence of 22 categorical variables was >0.99 with a mean absolute difference of 0.036. Overlapping principal component distributions confirmed the synthetic data's fidelity in representing inter-variable associations. The synthetic data demonstrated robust resilience against privacy intrusions with excellent scores for membership risk (i.e., determining if an individual had a record in the training dataset) and attribute risk (i.e., forecasting novel information about an individual). A proportional hazards model for time from ART initiation until death yielded similar results between synthetic and real data. While the synthetic data adequately ordered and calibrated survival across sites, they did not reproduce the early mortality seen in real data. These results illustrate both the potential and challenges with using synthetic data.

11:00 AM–11:20 AM Speaker: Jia Liang, St. Jude Children's Research Hospital

Information integration with varying coefficient model under high dimensional regression

Author(s): Jia Liang, St. Jude Children's Research Hospital; Chixiang Chen, University of Maryland; Shuo Chen, University of Maryland

Linear models are commonly used in clinical research to adjust for demographic effects such as age, weight, etc. Drawbacks are also significant since nonlinear effects are often ignored. Alternatively, the semi-varying coefficient model can be utilized to model both the nonlinear effect and the confounding structures. Together with high-dimensional structure in the linear covariates, this model has wide applications in economics, environmental science, and biomedical studies. This paper introduces a novel statistical inference framework that equips semi-varying coefficient model with high estimation efficiency by effectively synthesizing summary infor-

mation from external data into the main analysis. Such an integrative scheme is versatile in assimilating various reduced models from the external study while allowing heterogeneous covariate distribution from internal studies. The proposed method is proven theoretically valid and numerically convenient, and it enjoys a high-efficiency gain compared to classic methods in the semi-varying coefficient model.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S7: RECENT ADVANCES IN NON-/SEMI-PARAMETRIC MODELLING AND INFERENCE

Monday, June 17, 2024
10:00 AM–11:30 AM, Southern Ground A/B (Mezzanine Level)

Organizer: Tony Sit, The Chinese University of Hong Kong
Chair: Gongjun Xu, University of Michigan

10:00 AM–10:20 AM Speaker: Tony Sit, Chinese University of Hong Kong

Post selection inference for censored quantile regression

Author(s): Yu Guo, The Chinese University of Hong Kong; Tony Sit, The Chinese University of Hong Kong

This paper proposes a novel method for constructing confidence intervals for censored quantile regression in high-dimensional data settings where the number of covariates may significantly exceed the sample size. Building on the weighted loss function introduced by Wang and Wang (2009; Sinica), we apply an L1 penalisation and subsequently perform a debiasing process on the resulting estimate. The debiased estimator is shown to exhibit asymptotic normality, providing a robust basis for inference. Unlike existing research, our approach relaxes the global linearity condition to a local linearity condition near the quantile of interest, offering a more flexible and accurate model. This method is particularly advantageous when dealing with heteroskedastic effects or violations of global linearity. Simulation results demonstrate superior performance of our method in constructing confidence intervals.

10:20 AM–10:20 AM Speaker: Hok Kan Ling, Queen's University

Nonparametric likelihood ratio test for univariate shape-constrained densities

Author(s): Kwun Chuen Gary Chan, University of Washington; Hok Kan Ling, Queen's University; Chuan-Fa Tang, The University of Texas at Dallas; Sheung Chi Phillip Yam, The Chinese University of Hong Kong

We provide a comprehensive study of a nonparametric likelihood ratio test on whether a random sample follows a distribution in a prespecified class of shape-constrained densities. While the conventional definition of likelihood ratio is not well-defined for general nonparametric problems, we consider a working sub-class of alternative densities that leads to test statistics with desirable properties. Under the null, a scaled and centered version of the test statistic is asymptotic normal and distribution-free, which comes from the fact that the asymptotic dominant term under the null depends only on a function of spacings of transformed outcomes that are uniform distributed. The nonparametric maximum likelihood estimator (NPMLE) under the hypothesis class appears only in an average log-density ratio which often converges to zero at a faster rate than the asymptotic normal term under the null, while diverges in general test so that the test is consistent. The main technicality is to show these results for log-density ratio which requires a case-by-case analysis, including new results for k-monotone densities with unbounded support and completely monotone densities that are of independent interest. A bootstrap method by simulating from the NPMLE is shown to have the same limiting distribution as the test statistic.

10:40 AM–11:00 AM Speaker: Yue Xing, Michigan State University

Why do artificially generated data help adversarial robustness?
Author(s): Yue Xing, Michigan State University; Qifan Song, Purdue University; Guang Cheng, University of California, Los Angeles

> When people use generated/real unlabeled data with pseudolabels to improve adversarial robustness, the artificially generated data can improve adversarial training. We provide statistical insights to explain why the artificially generated data improve adversarial training. In particular, we study how the attack strength and the quality of the unlabeled data affect adversarial robustness in this framework. Our results show that with a high-quality unlabeled data generator, adversarial training can benefit greatly from this framework under large attack strength, while a poor generator can still help to some extent.

11:00 AM–11:20 AM Speaker: Chi Wing Chu, City University of Hong Kong

Competing risk quantile regression with time-dependent covariates
Author(s): Chi Wing Chu, City University of Hong Kong; Tony Sit, Chinese University of Hong Kong

> Quantile regression of the conditional quantile defined via the cumulative incidence function allows for the handling of dependent censoring from competing events and avoids the non-identifiability issue of the marginal distribution of the latent failure time. This work proposes a quantile regression model to accommodate right-censored competing risk data with time-dependent covariates. A recursive stepwise estimator of the regression parameter is derived via a martingale-based estimating equation and exploiting a conditional Kaplan-Meier estimator for time-dependent covariates. Asymptotic properties, including uniform consistency and weak convergence, are established via the underlying martingale structure. Monte Carlo simulations and numerical studies are presented to illustrate the numerical performance of the proposed estimator.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S8: UTILIZATION OF ADVANCED STATISTICAL DESIGNS AND ANALYSES IN CLINICAL TRIALS

Monday, June 17, 2024
10:00 AM–11:30 AM, Gold (Lower Level)

Organizer: Yimei Li, University of Pennsylvania
Chair: Yimei Li, University of Pennsylvania

10:00 AM–10:20 AM Speaker: Justine Shults, University of Pennsylvania

Accounting for the shutdown due to the COVID pandemic in a school and medical practice-based intervention: The West Philadelphia Asthma Care Implementation Study

Author(s): Justine Shults, Children's Hospital of Philadelphia and Perelman School of Medicine at the University of Pennsylvania; Chen Kenyon, Children's Hospital of Philadelphia and Perelman School of Medicine at the University of Pennsylvania; Andrea Apter, Perelman School of Medicine at the University of Pennsylvania; Julie Pappas, Westat and Children's Hospital of Philadelphia; Tyra Bryant-Stephens, Perelman School of Medicine at the University of Pennsylvania and Children's Hospital of Philadelphia

The West Philadelphia Asthma Care Implementation Program (WEPACC) was designed to improve asthma control in low-income children from communities that are disproportionately impacted by asthma and other serious pediatric health conditions. WEPACC is a randomized control trial that used a factorial design to compare usual care to interventions delivered by community healthcare workers in primary care, school alone, and combined primary care-school settings. However, the delivery of the WEPACC interventions was severely impacted by the SARS-CoV-2 pandemic. The shutdown of Philadelphia schools in Spring 2020 (roughly halfway through the study period) prevented the study team from delivering the school components as designed, while social distancing measures led to a striking decrease in asthma morbidity. Some participants completed all study visits prior to the shutdown for the pandemic, while others completed some (or all) of their study visits after the shutdown and therefore did not receive the full "dose" of the intervention as planned. In this presentation I will describe the challenges that we faced in the statistical analysis of data from this trial. I will also discuss and demonstrate the application of some recently published recommendations regarding how to properly modify the statistical analysis plan to account for unplanned interruptions in a clinical trial.

10:20 AM–10:20 AM Speaker: Chen Hu, Johns Hopkins University

On statistical inference of multiple competing risks in comparative clinical trials

Author(s): Jiyang Wen, Johns Hopkins University; Mei-Cheng Wang, Johns Hopkins University; Chen Hu, Johns Hopkins University

Competing risks data are commonly encountered in comparative clinical trials. Ignoring competing risks in survival analysis leads to biased risk estimates and improper conclusions. Often, one of the competing events is of primary interest and the rest competing events are handled as nuisances. These approaches can be inadequate when multiple competing events have important clinical interpretations and thus of equal interest. For example, in hospitalized critical care treatment trials, the outcomes are either death or discharge from hospital, which have completely different clinical implications and are of equal interest. In oncology trials, while com-

posite endpoints, such as disease-free survival, are used frequently, it is often concerned that novel interventions do not necessarily impact all components of a composite endpoint equally. We develop nonparamteric estimation and simultaneous inferential methods for multiple cumulative incidence functions (CIFs) and corresponding restricted mean times. Based on Monte Carlo simulations and a data analysis of a completed clinical trial, we demonstrate that the proposed method provides global insights of the treatment effects across multiple endpoints.

10:40 AM–11:00 AM Speaker: Yimei Li, University of Pennsylvania

Unlocking the potential: A systematic review of master protocols in pediatric trials
Author(s): Yimei Li, University of Pennsylvania

Master protocol is a novel clinical trial design that attempts to evaluate multiple experimental therapies in one or multiple indications, under one overarching protocol. The use of a master protocol holds the promise of increasing efficiency and enabling new approaches to statistical designs and analyses. However, the use of such novel designs in pediatric research is unclear. This study aims to provide a systematic review on the utilization of master protocols in pediatric clinical research. A systematic search was performed in September 2022 using two data sources (PubMed and ClinicalTrials.gov) and included studies conducted in the last 10 years. General study information was extracted such as study type, study status, therapeutic area, clinical trial phase. Study characteristics that are specific to pediatric studies (such as age of the participants and pediatric drug dosing) and important study design elements (such as number of test drug arms and whether randomization and/or concurrent control was used) were also collected. Our results suggested that master protocol studies are being used in pediatrics, with platform and basket trials more common than umbrella trials. The most experience is in oncology and early phase studies. There is a rise in the use starting in 2020, again largely in oncology. Application also emerged in COVID-19 trials. However, adoption of master in pediatric clinical research is still on a small scale and could be substantially expanded. Work is required to further understand the barriers in implementing pediatric master protocols, from setting up infrastructure to interpreting study findings.

11:00 AM–11:20 AM Speaker: Qi Wang, Duke University

Addressing population heterogeneity for HIV incidence estimation based on recency test
Author(s): Qi Wang, Duke University; Ann Duerr, Fred Hutchinson Cancer Center; Fei Gao, Fred Hutchinson Cancer Center

Cross-sectional HIV incidence estimation leverages recency test results to determine the HIV incidence of a population of interest, where recency test uses biomarker profiles to infer whether an HIV-positive individual was "recently" infected. This approach possesses an obvious advantage over the conventional cohort follow-up method since it avoids longitudinal follow-up and repeated HIV testing. In this manuscript, we consider the extension of cross-sectional incidence estimation to estimate the incidence of a different target population addressing potential population heterogeneity. We propose a general framework that incorporates two scenarios: one when the target population is a subset of the population with cross-sectional recency testing data, and the other with an external target population. In addition, we propose estimators to incorporate HIV subtypes, a special type of covariate that modifies the properties of recency test, into our framework. Through extensive simulation studies and a data application, we demonstrate the

performance of the proposed methods. We conclude with a discussion on sensitivity analysis and future work to improve our framework.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S9: STATISTICAL CHALLENGES OF MODELING PATIENT-REPORTED OUTCOME MEASURES IN REAL-WORLD EVIDENCE STUDIES

Monday, June 17, 2024
10:00 AM–11:30 AM, Melody (Lobby Level)

Organizer: Xiaofeng Wang, Cleveland Clinic
Chair: Yaomin Xu, Vanderbilt University Medical Center

10:00 AM–10:20 AM Speaker: Edward Ip, Wake Forest School of Medicine

Modeling patient reported outcomes that are only meaningful at one end of the distribution

Author(s): Edward Ip, Wake Forest School of Medicine

Patient Reported Outcome (PRO) data often exist as ordinal or categorical responses on a collection of items or survey questions answered by a patient. Unlike clinician rated outcomes, PROs reflect patients' perspectives about treatment effects, and represent outcomes that are of upmost importance to patients and families. Derived from the traditional measurement paradigm, PROs are designed to assess bipolar traits, which hold significance at both ends of the scale. However, certain constructs in the medical domain, such as depression, alcoholism, and to a certain extent adverse events, present as unipolar traits, where the trait is meaningful only at one end of the distribution. For instance, a low score indicates the absence of a quality (e.g., not alcoholic) rather than a relatively low score compared to others with that quality (i.e., less alcoholic). This implies a qualitative difference between the two groups, making it inappropriate to place them on the same scale. Additionally, certain items within PROs, such as suicidal ideation in a depression inventory, possess higher discriminatory power than others (e.g., feeling blue). Endorsement of a single discriminatory item thus could signal problem, which may not be captured by traditional scaling method such as sum score. Conventional statistical methods like inflated zero may not be effective for modeling unipolar traits, as a non-zero low score, in the case of depression, can also indicate the absence of the condition. One practical approach is to establish thresholds to delineate different categories of the condition. However, this threshold-based approach often necessitates extensive expert input and consensus. This presentation aims to address these challenges by exploring various methodologies including latent variable modeling and partially ordered set theory to construct measurement models for unipolar traits.

10:20 AM–10:40 AM Speaker: Joseph Cappelleri, Pfizer, Inc.

Reflective vs. formative measurement models for psychometric validation

Author(s): Joseph Cappelleri, Pfizer Inc.

In psychometric validation, a measurement model involves the relationship between the latent variables (domains) and their observed items or variables or questions (indicators). In a reflective model, a latent variable drives the indicators (known as effect or reflective indicators), which includes a majority of measurement work in the social and health sciences. In a formative model, on the other hand, the indicators (also called causal or formative indicators) influence its latent concept; therefore, a change in an indicator affects a change in the concept or construct (rather than the reverse). The implications of these two types of models—reflective and formative—for psychometric validation are discussed conceptually and with a real-life example and related simulation study.

10:40 AM–11:00 AM Speaker: Bin Wang, US Food and Drug Administration

Evaluating the effectiveness by employing multi-state models based on patient-reported outcomes from a randomized-controlled clinical trial

Author(s): Bin Wang, US Food and Drug Administration (FDA); Xuefueng Li, FDA; Yu Zhao, FDA; Saryet Kucukemiroglu, FDA

In clinical studies, the inherent heterogeneity among patient populations is a common challenge, compounded by varying treatment effects. Additionally, the ceiling or floor effects due to extreme response scales and the non-ignorable between-subject variation in patient-reported outcomes can significantly influence effectiveness evaluations. Our proposed approach involves employing multi-state models to systematically assess both the heterogeneity within patient cohorts and the effectiveness of treatments. This will be achieved through a comprehensive analysis of longitudinal patient-reported outcomes, shedding light on nuanced variations and providing a robust framework for evaluating treatment effects.

11:00 AM–11:20 AM Discussant: Xiaofeng Wang, Cleveland Clinic

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S10: INNOVATIVE APPLICATIONS IN CLINICAL TRIALS

Monday, June 17, 2024
10:00 AM–11:30 AM, Green Room (Lobby Level)

Organizer: Zhe Qu, Servier BioInnovation
Chair: Qian Meng, Servier BioInnovation

10:00 AM–10:20 AM Speaker: Daoyuan Shi, BlueRock Therapeutics

Statistical transplantation and erythropoiesis model in clinical trials

Author(s): Daoyuan Shi, BlueRock Therapeutics

Cell and gene therapies are at the forefront of innovation, transforming how we treat and potentially cure certain diseases. For many gene and cell therapies, edited cells are infused back into the human body to take effect. It takes time for the edited cells to grow and achieve equilibrium with the host's original unedited cells. The time to achieve equilibrium and the proportion of edited cells in that equilibrium are two important questions for drug developers. Previously, clinical teams relied on the literature and their experience to make estimations. We have developed an innovative statistical transplantation and erythropoiesis model using an iterative approach to mimic the erythropoiesis process. With different simulation parameters, the model could provide guidance on the selection of clinical endpoints, the setting of manufacturing specifications, and treatment dosage. The model could also be extended to other therapeutic areas.

10:20 AM–10:40 AM Speaker: Yimei Li, St. Jude Children's Research Hospital

Group sequential multi-arm multi-stage trial design with treatment selection

Author(s): Jianrong Wu, University of New Mexico; Yimei Li, St Jude; Liang Zhu,

A multi-arm trial allows simultaneous comparison of multiple experimental treatments with common control and provides a substantial efficiency advantage compared to the traditional randomized controlled trial. Many novel multi-arm multi-stage (MAMS) clinical trial designs have been proposed. However, a major hurdle to adopting the group sequential MAMS routinely is the computational effort of obtaining the total sample size and sequential stopping boundaries. In this paper, we develop a group sequential MAMS trial design based on the sequential conditional probability ratio test. The proposed method provides analytical solutions for futility and efficacy boundaries to an arbitrary number of stages and arms. Thus, it avoids complicated computational effort for the methods proposed by Magirr et al. Simulation results showed that the proposed method has several advantages compared to the methods implemented in the R package MAMS by Magirr et al.

10:40 AM–11:00 AM Speaker: Zhaowei Hua, Servier BioInnovation

Application of artificial intelligence in clinical trial supply chain management

Author(s): Zhaowei Hua, Servier BioInnovation; Jincheng Pang, Washington University in St. Louis; Hong Yan, Servier BioInnovation

Emerging pivotal challenges from the intricate landscape of drug supply chain management can

potentially offset the benefit in applying innovative adaptive designs in clinical trials. The challenges include the uncertainty of maximum drug supply needed, shifting of supply requirement, cost control, and high-dimension factors impacting the decision of drug resupply. To address these issues, we designed an optimization digital tool tailored for the efficient management of drug supply in clinical trials. The tool optimizes drug supply strategies in a sequential manner throughout the trial, leveraging real-time data and statistical simulations to make informed decisions. Real world scenarios are integrated into the framework as pragmatic assumptions and setup. Statistical simulations are applied to optimize drug supply strategy in pre-study planning stage as well as during study monitoring stage. Drug supply optimization is realized by minimizing the total cost using real-time data from the study over time. An artificial intelligence model particle swarm optimization algorithm is applied to perform optimization, where feature extraction is implemented to reduce dimensionality and computational cost.

11:00 AM–11:20 AM Speaker: Cong Li, Takeda Pharmaceuticals

Application of marginal structural models in oncology randomized controlled trials (RCTs) to adjust for subsequent therapy effect on overall survival

Author(s): Jing Xu, Takeda Development Center Americas; Camden Bay, Takeda Development Center Americas; Bingxia Wang, Takeda Development Center Americas; Guohui Liu, Takeda Development Center Americas; Cong Li, Takeda Development Center Americas

In oncology randomized controlled trials (RCTs), patients are usually permitted to take alternative treatments after disease progression because of ethical considerations. In such cases, the effect of active intervention on overall survival (OS) is confounded by the effect of subsequent therapies. In this presentation, we first describe the application of marginal structural Cox proportional hazard model (referred to MSMs) in adjusting for such confounding effect in OS, which has been promoted in the statistics community recently. Then, we focus on a special case of crossover design, where only patients in the control arm are allowed to crossover to the active treatment arm, in which the problem of structural non-positivity may prevent a direct use of MSM. We propose a two-step approach to solve this problem and allows for MSMs to be used in this scenario. We show the validity of the proposed approach through a simulation study and also illustrate its use through a case study.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

P1: TRAINING THE NEXT GENERATION OF BIOPHARMACEUTICAL STATISTICIANS

Monday, June 17, 2024
1:00 PM–2:30 PM, Symphony 1(Lobby Level)

Organized by Erik Bloomquist, Senior Principal Scientist at Merck, and moderated by Chenguang Wang, Senior Director at Regeneron, this panel of leading biopharmaceutical statisticians will share their observations and insights on helping the next generation meet the field's new and ongoing challenges. The panelists are:

- Samuel Wu, Associate Chair, Department of Biostatistics, University of Florida
- Naitee Ting, Director, Department of Biostatistics & Data Sciences, Boehringer-Ingelheim
- Jing-ou Liu, Vice President, Biostatistics & Data Management, Regeneron
- Brian Rini, Chief of Clinical Trials, Vanderbilt-Ingram Cancer Center
- Jing Huang, Senior Vice President, Bioinformatics and Data Science, Veracyte

S11: INNOVATIVE METHODS TO ADDRESS EMERGING CHALLENGES IN EVENT TIME DATA ANALYSIS

Monday, June 17, 2024
1:00 PM–2:30 PM, Blackbird A (Mezzanine Level)

Organizer: Donglin Zeng, University of Michigan
Chair: Yuhao Deng, University of Michigan

1:00 PM–1:20 PM Speaker: Yi Li, University of Michigan

Deep learning of partially linear Cox models

Author(s): Yi Li, University of Michigan

Lung cancer is a leading cause of cancer mortality globally, highlighting the importance of understanding its mortality risks to design effective patient-centered therapies. The National Lung Screening Trial (NLST) employed computed tomography texture analysis, which provides objective measurements of texture patterns on CT scans, to quantify the mortality risks of lung cancer patients. Partially linear Cox models have gained popularity for survival analysis by dissecting the hazard function into parametric and nonparametric components, allowing for the effective incorporation of both well-established risk factors (such as age and clinical variables) and emerging risk factors (e.g., image features) within a unified framework. However, when the dimension of parametric components exceeds the sample size, the task of model fitting becomes formidable, while nonparametric modeling grapples with the curse of dimensionality. We propose a novel Penalized Deep Partially Linear Cox Model (Penalized DPLC), which incorporates the SCAD penalty to select important texture features and employs a deep neural network to estimate the nonparametric component of the model. We prove the convergence and asymptotic properties of the estimator and compare it to other methods through extensive simulation studies, evaluating its performance in risk prediction and feature selection. The proposed method is applied to the NLST study dataset to uncover the effects of key clinical and imaging risk factors on patients' survival. Our findings provide valuable insights into the relationship between these factors and survival outcomes.

1:20 PM–1:40 PM Speaker: Lang Zeng, University of Pittsburgh

Mini-batch stochastic gradient descent for Cox regression and neural network: Theoretical foundation and practice guidance

Author(s): Lang Zeng, University of Pittsburgh; Weijing Tang, Carnegie Mellon University; Ying Ding, University of Pittsburgh

Large-scale data presents challenges to the typical optimization of Cox Proportional Hazards (CoxPH) models, which computes the gradients over the entire dataset. The mini-batch stochastic gradient descent (SGD) algorithm is a scalable solution for optimizing CoxPH models in large-scale data environments. Although the numerical success of SGD in CoxPH models and their neural network-based variants is well documented, the underlying reasons for its effectiveness remain less understood. In this work, we establish theoretical foundations to justify the application of mini-batch SGD in CoxPH models. Furthermore, we provide both numerical and theoretical evidence demonstrating the improvement in statistical efficiency when doubling the mini-batch size in CoxPH regression. Additionally, we demonstrate the failure of the widely

used linear learning rate scaling rule in the CoxPH neural network. These two phenomena are not typically observed in traditional mini-batch SGD optimizations. We also offer practical guidance on the learning rate scaling for mini-batch SGD in the CoxPH neural network. We conclude that the non-decomposability of the CoxPH loss function contributes to these unique characteristics. Our results have broader implications and can be generalized to other mini-batch SGD optimizations involving non-decomposable loss functions, such as learning-to-rank and contrastive learning.

1:40 PM–2:00 PM Speaker: Peijun Sang, University of Waterloo

Functional principal component analysis with informative observation times

Author(s): Peijun Sang, University of Waterloo; Dehan Kong, University of Toronto; Shu Yang, North Carolina State University

Functional principal component analysis has been shown to be invaluable for revealing variation modes of longitudinal outcomes, which serves as important building blocks for forecasting and model building. Decades of research have advanced methods for functional principal component analysis often assuming independence between the observation times and longitudinal outcomes. Yet such assumptions are fragile in real-world settings where observation times may be driven by outcome-related reasons. Rather than ignoring the informative observation time process, we explicitly model the observational times by a counting process dependent on time-varying prognostic factors. Identification of the mean, covariance function, and functional principal components ensues via inverse intensity weighting. We propose using weighted penalized splines for estimation and establish consistency and convergence rates for the weighted estimators. Simulation studies demonstrate that the proposed estimators are substantially more accurate than the existing ones in the presence of a correlation between the observation time process and the longitudinal outcome process. We further examine the finite-sample performance of the proposed method using the Acute Infection and Early Disease Research Program study.

2:00 PM–2:20 PM Speaker: Weijing Tang, Carnegie Mellon University

Recurrent event analysis with ordinary differential equations

Author(s): Weijing Tang, Carnegie Mellon University

This talk presents a general framework for the analysis of recurrent event data. The framework accommodates a wide range of semi-parametric recurrent event models by modeling the conditional mean function associated with the recurrent event process as the solution of an Ordinary Differential Equation (ODE), including both non-homogeneous Poisson and non-Poisson processes. We propose a Sieve Maximum Pseudo-Likelihood Estimation method, which employs the Non-Homogeneous Poisson Process (NHPP) as the working model. We also establish the consistency, convergence rate, and asymptotic normality of the proposed estimator, achieving semi-parametric efficiency when the NHPP working model is correct.

Additionally, we implement an efficient resampling method to estimate the asymptotic variance. To validate our proposed method's statistical efficiency and computational scalability, we conduct extensive numerical studies, including various simulation settings and a real-world dataset focused on analyzing risk factors for hospital readmission. This is joint work with Bo Meng, Gongjun Xu, and Ji Zhu.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S12: MACHINE LEARNING METHODS FOR COMPLEX DATA ANALYSIS AND MODELING

Monday, June 17, 2024
1:00 PM–2:30 PM, Lyric (Lobby Level)

Organizer: Baiming Zou, University of North Carolina at Chapel Hill
Chair: Yaomin Xu, Vanderbilt University Medical Center

1:00 PM–1:20 PM Speaker: Andrei Rodin, City of Hope

A new universal scoring criterion for probabilistic graphical models that enables automated multi-scale network inference and inter-network comparisons

Author(s): Grigoriy Gogoshin, City of Hope National Medical Center / Beckman Research Institute; Sergio Branciamore, City of Hope National Medical Center / Beckman Research Institute; Andrei Rodin, City of Hope National Medical Center / Beckman Research Institute

One of the principal limitations of probabilistic graphical models, such as Bayesian networks (BNs), is the non-commensurate and unbound nature of their model scoring functions (typically, BIC/MDL or AIC). This makes it difficult to directly compare BNs generated even from the homologous datasets (e.g., sharing the same feature set), let alone multiscale BNs originating from different domains. This, in turn, compromises the adoption of the probabilistic graphical models for many otherwise well-suited tasks in diverse biomedical research areas. A typical example would be comparing the multiscale/multimodal networks for cases vs. controls, or responders to a particular therapy vs. non-responders, or "before" vs. "after" timeslices in a dynamic series. In this talk, we present a novel model scoring criterion, Minimum Uncertainty (MU), that is commensurate and bound to a universal (0-1) scale. We detail our BN analysis pipeline and show a number of example applications. Finally, we discuss the relative pros and cons of probabilistic graphical models and graph neural networks (GNNs) in the context of next-generation cancer and oncology research.

1:20 PM–1:40 PM Speaker: George Tseng, University of Pittsburgh

High-dimensional unsupervised machine learning with multi-facet structure and outcome guidance in omics disease subtyping applications

Author(s): George Tseng, University of Pittsburgh

High-dimensional omics data often contain intricate and multifaceted information, resulting in the coexistence of multiple plausible sample partitions based on different subsets of selected features. Conventional clustering methods typically yield only one clustering solution, limiting their capacity to fully capture all facets of cluster structures in high-dimensional data. To address this challenge, we propose a model-based Multi-Facet Clustering (MFClust) method based on a mixture of Gaussian mixture models, where the former mixture achieves facet assignment for gene features and the latter mixture determines cluster assignment of samples. We demonstrate superior facet and cluster assignment accuracy of MFClust through simulation studies. The proposed method is applied to three transcriptomic applications from postmortem brain and lung disease studies. The result captures multi-facet clustering structures associated with critical clinical variables and provides intriguing biological insights for further hypothesis generation and discovery.

1:40 PM–2:00 PM Speaker: Yaomin Xu, Vanderbilt University Medical Center

Learning disease multimorbidity patterns across multiple EHR systems

Author(s): Yaomin Xu, Vanderbilt University Medical Center; Nick Strayer, Posit; Siwei Zhang, Vanderbilt University; Tess Vessels, Vanderbilt University; Douglas Ruderfer, Vanderbilt University Medical Center

Multimorbidity, where multiple health conditions co-exist non-randomly within an individual, is a growing challenge for healthcare and society. Understanding multimorbidity patterns can lead to better prevention, treatments, and personalized care. The advent of electronic health record (EHR) systems provides a vast trove of data for studying real-world patient health dynamics. However, concerns about the primary design of EHRs for billing and administration raise questions about the consistency and reproducibility of EHR-based research. In this study, we used the International Classification of Diseases (ICD) codes to analyze disease comorbidity patterns and employed network modeling to examine multimorbidity across two major EHR systems. Our findings revealed highly correlated multimorbidity patterns across HER systems, with graph-theoretic analysis confirming the consistency of the multimorbidity networks at local (nodes and edges), global (network statistics) and meso (neighboring connection structures) scales. This result offered new insights for developing an efficient framework to analyze and compare complex structures within the multimorbidity network. Our case study demonstrated that identifying subgraphs within multimorbidity networks is an effective method for detecting disease condition clusters, and, supported by graph spectral characteristics of the multimorbidity networks, we developed a complete online network clustering algorithm as an efficient approach to identify those clusters. To facilitate access to these complex datasets and promote further discovery research and hypothesis generation, we have developed a suite of interactive visualization tools for complex online data analysis leveraging data from multiple EHR/Biobank data sources. These tools are open source, available to the public, and are designed to enable researchers to intuitively explore the complex disease relationships within the multimorbidity networks, thereby enhancing our collective understanding and fostering the development of novel precision medicine solutions in the context of multimorbidities.

2:00 PM–2:20 PM Speaker: Baiming Zou, University of North Carolina at Chapel Hill

A deep neural network two-part model and feature importance test for semi-continuous data

Author(s): Baiming Zou, University of North Carolina at Chapel Hill

Semi-continuous data frequently arise in clinical practice. For example, while many surgical patients still suffer from varying degrees of acute postoperative pain (POP) sometime after surgery (i.e., POP score >0), others experience none (i.e., POP score=0), indicating the existence of two distinct data processes at play. Existing parametric or semi-parametric two-part modeling methods for this type of semi-continuous data can fail to appropriately model the two underlying data processes as such methods rely heavily on (generalized) linear additive assumptions. However, many factors may interact to jointly influence the experience of POP non-additively and non-linearly. Motivated by this challenge and inspired by the flexibility of deep neural networks (DNN) to accurately approximate complex functions universally, we derive a DNN-based two-part model by adapting the conventional DNN methods with two additional components: a bootstrapping procedure along with a filtering algorithm to boost the stability of the conventional DNN, an approach we denote as sDNN. To improve the interpretability

and transparency of sDNN, we further derive a feature importance testing procedure to identify important features associated with the outcome measurements of the two data processes, denoting this approach fsDNN. We show that fsDNN not only offers a statistical inference procedure for each feature under complex association but also that using the identified features can further improve the predictive performance of sDNN. The proposed sDNN- and fsDNN-based two-part models are applied to the analysis of real data from a POP study, in which application they clearly demonstrate advantages over the existing parametric and semi-parametric two-part models. Further, we conduct extensive numerical studies and draw comparisons with other machine learning methods to demonstrate that sDNN and fsDNN consistently outperform the existing two-part models and frequently used machine learning methods regardless of the data complexity.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S13: NOVEL STATISTICAL METHODS FOR IMAGING DATA ANALYSIS AND BEYOND

Monday, June 17, 2024
1:00 PM–2:30 PM, Ocean Way (Mezzanine Level)

Organizer: Panpan Zhang, Vanderbilt University Medical Center
Chair: Panpan Zhang, Vanderbilt University Medical Center

1:00 PM–1:20 PM Speaker: Andrew Chen, Medical University of South Carolina

Structure-function gradients along the brain cortex

Author(s): Andrew Chen, Medical University of South Carolina

Recent methodological advances allow examining the topological organization of the brain cortex and deriving gradients of organization. These gradients are consistent with seminal research on brain functional organization, well-studied neurodevelopmental trajectories, and measures of association between structural and functional measures. This structure-function relationship has attracted particular interest due to its association with known biological changes. However, gradients have generally been estimated solely through functional imaging and have yet to capture structural trends in the brain. A novel method is proposed for the derivation of structure-function gradients from both functional magnetic resonance imaging and diffusion tensor imaging. Application to the Philadelphia Neurodevelopmental Cohort reveals that these novel gradients reflect well-known brain organization while suggesting novel cortical patterns. The method extends more generally to any multimodal brain measurements and potential extensions and statistical frameworks are explored for downstream analyses.

1:20 PM–1:40 PM Speaker: Moo Chung, University of Wisconsin-Madison

Unraveling dynamic brain networks: A topological data analysis framework for state space estimation

Author(s): Moo Chung, University of Wisconsin-Madison

We present a novel topological data analysis (TDA) framework for modeling state spaces in dynamic functional brain networks. Core principles of TDA are explored, particularly highlighting how topological distances can be harnessed to cluster time-evolving brain networks. The approach is centered on the use of Wasserstein metrics to measure distances between 0D and 1D topological features, which significantly enhances the ability to track the brain's connectivity patterns over time. Our approach outperforms traditional methods such as k-means and hierarchical clustering by capturing the hidden topological patterns in the data, which are ignored in existing methods. The analysis reveals that resting-state brain networks oscillate primarily between two major states, with an additional noisy intermediate state. These fluctuations are linked to the complex sulcal and gyral cortical patterns, where significant differences in functional and structural connectivity are observed. The variability in connectivity introduces a dynamic gradient across the network, influencing the overall dynamics of brain connectivity changes. The talk is partially based on arXiv:2201.00087 (PLOS Computational Biology).

1:40 PM–2:00 PM Speaker: Shan Yu, University of Virginia

Distributed learning for large-scale longitudinal image-on-scalar regression

Author(s): Hyunjae Cho, University of Virginia; Shan Yu, University of Virginia

Neuroimaging studies are becoming increasingly important in medical fields as the prevalence of Alzheimer's disease. Our study is motivated by investigating brain activities and exploring the relationship between varying scalar predictors and different brain regions over time through longitudinal image-on-scalar regression. However, the difficulty lies in brain regions' complexity, the sparsity of time measurement for each subject, and the massive computational costs associated with analyzing large scalar imaging data. To address these issues, we proposed an individual growth path model based on image-on-scalar regression in the nonparametric functional data analysis framework. The bivariate penalized spline over triangulation (BPST) method is used to handle the irregular domain of brain images for estimating the coefficient function. We propose a novel approach to parallel computing that utilizes Hilbert space-filling curve-based domain decomposition on BPST (HBDB) to reduce the computational time. The proposed nonparametric varying coefficient functions in both BPST and HBDB methods are proven to be asymptotically normal and consistent under some regularity conditions. The proposed methods are evaluated through extensive simulation studies and analyses of studies in the Alzheimer's Disease Neuroimaging Initiative.

2:00 PM–2:20 PM Speaker: Simiao Gao, Yale University

Statistical methods for analyzing lifespan biomarkers: Application to HCP aging data

Author(s): Simiao Gao, Yale University; Yifei Sun, Columbia University; Yize Zhao, Yale University

Biased sampling design is often encountered in research problems. Suppose we are interested in a midlife and aging population (e.g., $\geq$ age 35) that have not encountered pathology or neurodegeneration onset. In this case, only partial individuals who are event free at baseline and are qualified to enter the study. Thus, the study sample is not a random sample from the target population because it favors individuals with longer event times. We propose nonparametric and semiparametric methods to estimate a composite trait/endophenotype on both event onset and marker dynamics, defined as the area under the marker trajectory up to the pathology or neurodegeneration onset. Following that, we then conduct extensive simulations to evaluate the accuracy and efficacy of our proposed model. Finally, we then apply our model to the HCP Aging data set to discover the effects of sex and sleeping quality index on the sub-networks of brain functional connectivity.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S14: RECENT ADVANCEMENTS FOR STATISTICAL LEARNING ON COMPLEX DATA

Monday, June 17, 2024
1:00 PM–2:30 PM, Platinum (Lower Level)

Organizer: Zhengwu Zhang, University of North Carolina at Chapel Hill
Chair: Zhengwu Zhang, University of North Carolina at Chapel Hill

1:00 PM–1:20 PM Speaker: Tingting Zhang, University of Pittsburgh

Analysis of functional connectivity changes from childhood to old age: A study using HCP-D, HCP-YA, and HCP-A datasets

Author(s): Yaotian Wang, Emory University; Shuoran Li, University of Pittsburgh; Jie He, University of Pittsburgh; Lingyi Peng, University of Pittsburgh; Dana Tudorascu, University of Pittsburgh; Lauren Schaeffer, University of Pittsburgh; Stacey Sukoff Rizzo, University of Pittsburgh; Gregory Carter, Jackson Laboratory; Afonso Silva, University of Pittsburgh; Tingting Zhang, University of Pittsburgh

We present a new clustering-enabled regression approach designed to investigate how whole-brain functional connectivity (FC) in healthy subjects changes from childhood to old age. By applying this method to aggregated fMRI data from three Human Connectome Project studies, we identify clusters of brain regions with identical patterns of FC changes over time and map out these FC trajectories across the identified clusters. Our findings reveal that age affects FC in a varied manner across different brain regions. Most brain connections experience minimal yet statistically significant FC changes with age. Only a tiny proportion of connections exhibit substantial age-related changes in FC. Among these connections, FC between brain regions in the same functional network tends to decrease over time, while FC between regions in different networks demonstrates diverse patterns of age-related changes, underscoring the intricate nature of brain aging processes. Moreover, our research uncovers sex-specific trends in FC changes; while the average FC is comparable between females and males in childhood, it becomes increasingly different with aging. Elderly females show much higher FC within the default mode network and in certain between-network connections of the somatomotor network, whereas elderly males display higher FC across multiple brain networks. Furthermore, our study suggests that the relationship between cognitive behavior and FC is nuanced, being most influenced by age and sex during childhood, less influenced in older adults, and to the least extent in young adults.

1:20 PM–1:40 PM Speaker: Fei Jiang, University of California, San Francisco

On high dimensional Poisson models with measurement error: Hypothesis testing for nonlinear nonconvex optimization

Author(s): Fei Jiang, University of California, San Francisco

We study estimation and testing in the Poisson regression model with noisy high dimensional covariates, which has wide applications in analyzing noisy big data. Correcting for the estimation bias due to the covariate noise leads to a non-convex target function to minimize. Treating the high dimensional issue further leads us to augment an amenable penalty term to the target function. We propose to estimate the regression parameter through minimizing the penalized target function. We derive the L1 and L2 convergence rates of the estimator and prove the

variable selection consistency. We further establish the asymptotic normality of any subset of the parameters, where the subset can have infinitely many components as long as its cardinality grows sufficiently slow. We develop Wald and score tests based on the asymptotic normality of the estimator, which permits testing of linear functions of the members if the subset. We examine the finite sample performance of the proposed tests by extensive simulation. Finally, the proposed method is successfully applied to the Alzheimer's Disease Neuroimaging Initiative study, which motivated this work initially.

1:40 PM–2:00 PM Speaker: Fei Zou, University of North Carolina at Chapel Hill

Advanced machine learning models for genetics and genomics data

Author(s): Fei Zou, University of North Carolina at Chapel Hill

Deep Neural Network has become one of the most popular machine learning algorithms in biomedical research due to its high flexibility in approximating complex functions. In this talk, we will present a deep neural network based algorithm for personalized drug response prediction and a computational pipeline for identifying antigens that stimulate immune response (i.e., peptide presentation). In addition, to identify important biomarkers and to improve model interpretations of these seemingly "black-box" models, an empirical feature importance scoring technique will be discussed.

2:00 PM–2:20 PM Speaker: William Consagra, Harvard Medical School

Continuous and atlas-free analysis of brain structural connectivity

Author(s): William Consagra, Harvard Medical School; Martin Cole, University of Rochester Medical Center; Xing Qiu, University of Rochester Medical Center; Zhengwu Zhang, University of North Carolina at Chapel Hill

Brain structural networks are often represented as discrete adjacency matrices, with elements summarizing the connectivity between pairs of regions of interest (ROIs). These ROIs are typically determined a-priori using a brain atlas. The choice of atlas is often arbitrary and can lead to a loss of important connectivity information at the sub-ROI level. This work introduces an atlas-free framework that overcomes these issues by modeling brain connectivity using smooth random functions. In particular, we assume that the observed pattern of white matter fiber tract endpoints is driven by a latent random function defined over a product manifold domain. To facilitate statistical analysis of these high dimensional functional data objects, we develop a novel algorithm to construct a data-driven reduced-rank function space that offers a desirable trade-off between computational complexity and flexibility. Using real data from the Human Connectome Project, we show that our method outperforms state-of-the-art approaches that use the traditional atlas-based structural connectivity representation on a variety of connectivity analysis tasks. We further demonstrate how our method can be used to detect localized regions and connectivity patterns associated with group differences.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S15: INNOVATIVE AND APPLIED METHODS IN DESIGN AND ANALYSIS OF ONCOLOGY TRIALS ON THE TIME-TO-EVENT ENDPOINTS

Monday, June 17, 2024
1:00 PM–2:30 PM, Sound Emporium A/B (Mezzanine Level)

Organizer: Huan Cheng, BeiGene
Chair: Huan Cheng, BeiGene

1:00 PM–1:20 PM Speaker: Huan Cheng, BeiGene

A general approach for sample size calculation with non-proportional hazards and cure rate

Author(s): Huan Cheng, BeiGene

With the ongoing advancements in cancer treatment, a subset of patients now experiences long-term survival or even achieves a cure in certain cancer types. Additionally, non-proportional hazards such as delayed treatment effects and crossing hazards are commonly observed in immuno-oncology clinical trials. To address these challenges, various cure models have been proposed to integrate the cure rate into trial design and accommodate delayed treatment effects. In this article, we introduce a unified approach for calculating sample sizes, taking into account different cure rate models and non-proportional hazards. Our approach supports both the traditional weighted logrank test and the Maxcombo test, which demonstrates robust performance under nonproportional hazards. Furthermore, we assess the accuracy of our sample size estimation through Monte Carlo simulation across diverse scenarios and compare our method with existing approaches. Several illustrative examples are provided to demonstrate the proposed methodology.

1:20 PM–1:40 PM Speaker: Jianqi Zhang, Amgen Inc.

PubPredict: Prediction of progression free and overall survival in oncology trials leveraging publications and early efficacies

Author(s): Jianqi Zhang, Amgen Inc.; Junyi Zhou, Amgen Inc.; Erik Rasmussen, Amgen Inc.

In oncology/hematology early phase clinical trials, early efficacies were often observed in terms of response rate, depth, timing and duration. However, the true clinical benefits that eventually support registrational purpose of a phase 3 study are progression-free survival (PFS) and/or overall survival (OS), the follow-up of which are typically not long enough in early phase trials. This imposes challenges in strategies and decision making in drug development. In this work, we tackle the challenge by capturing the connection between early and late efficacy endpoints through a continuous Markov chain (CMC) process. We developed a parameterization algorithm for the transition intensity matrix of a CMC model, that would govern a CMC process that, when transformed to time-to-event data, would reflect published aggregate-level summary statistics. Then, transition intensity matrix could be modified to reflect various scenarios of response rate, depth, timing and duration, which allows predictions of median PFS and/or OS. A R shiny application named PubPredict is built for interactive communications with cross-functional colleagues. The tool implements the algorithm described above and allows customized features including multiple response levels, treatment crossover and varying follow-up durations. This toolset has been applied to advise phase 3 trial design.

1:40 PM–2:00 PM Speaker: Anjun Cao, Bayer U.S. LLC

Survival analysis for time-to-event endpoint with interval-censored data in clinical trials

Author(s): Anjun Cao, Bayer U.S. LLC

In survival analysis, we often encounter interval-censored time-to-event data. For example, in oncology studies, progressive-free survival (PFS) is based on the tumor assessments at scheduled visits. The exact time of PFS events is unknown. My research includes two parts. first, I explored the impact of frequency of assessments on the observed median improvement between the treatments. My simulation results show the frequency of assessments with the assessment intervals 50%-70% of the true median improvement provides high probability to observe median improvement larger than the assessment interval. Second, I compared the analysis results using right-censored statistical method and interval-censored statistical method. There is no conclusion to draw that one is superior to the other. The choice is based on individual case. The observed median time to event and median improvement truly depend on the combination factors of distributions of time to event for both treatments, assessment interval, distribution of time to dropout, and possible missing assessments immediately prior to the event. It is helpful to add an interval-censored analysis as a sensitivity analysis in addition to the primary analysis using right-censored statistical method for the interval-censored data.

2:00 PM–2:20 PM Discussant: Qingxia Chen, Vanderbilt University Medical Center

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S16: RECENT DEVELOPMENTS IN STATISTICS AND MACHINE LEARNING FOR ELECTRONIC HEALTH RECORDS DATA AND PRECISION MEDICINE

Monday, June 17, 2024
1:00 PM–2:30 PM, Southern Ground A/B (Mezzanine Level)

Organizer: Zuoheng Wang, Yale University
Chair: Zuoheng Wang, Yale University

1:00 PM–1:20 PM Speaker: Xiayuan Huang, Yale University

Enhancing patient representation learning from electronic health records through predicted family relations

Author(s): Xiayuan Huang, Yale University; Johann de Jong, Boehringer Ingelheim; Zuoheng Wang, Yale University

Artificial intelligence and machine learning are powerful tools for analyzing electronic health records (EHRs) in healthcare research. Despite the recognized importance of family health history, patients are often treated as independent samples in traditional analyses, overlooking family relations. To address this gap, we present ALIGATEHR, which models predicted family relations in a graph attention network integrated with a medical ontology. Taking disease risk prediction as a use case, we first demonstrate that explicitly modeling family relations significantly improves predictions across the disease spectrum. We then show how ALIGATEHR's attention mechanism successfully captures genetic aspects of diseases using only EHR diagnosis data. Finally, we use ALIGATHER to successfully distinguish the two main inflammatory bowel disease subtypes (Crohn's disease and ulcerative colitis). Our results highlight that family relations should not be overlooked in EHR research and illustrate ALIGATEHR's great potential for improving patient representation learning for predictive and descriptive modeling of EHRs.

1:20 PM–1:40 PM Speaker: Weidong Ma, University of Pennsylvania

A semiparametric method for addressing under-diagnosis using electronic health record data

Author(s): Weidong Ma, University of Pennsylvania Perelman School of Medicine; Jinbo Chen, University of Pennsylvania Perelman School of Medicine

Under-diagnosis, which occurs when patients live with a disease condition without receiving a diagnosis, prevents patients from obtaining suitable treatments and preventive strategies. Electronic Health Records (EHRs) contain a wealth of patient information and offer a unique opportunity to identify under-diagnosis, as diagnosed and under-diagnosed patients may exhibit similarities in their EHR profiles, which differ from those of disease-free patients. However, this opportunity to date has not been fully exploited due to the "positive-unlabeled" structure of EHR data, where under-diagnosed patients are mixed together with a large number of disease-free patients. To address this challenge, we develop a novel statistical approach based on importance weighting method to enable unbiased assessment of the risk that a patient has the disease condition, where EHR data is supplemented with a small number of additional disease labels acquired through targeted screening for patients who have not received a diagnosis. The performance of the proposed method is studied via characterization of asymptotic properties

and extensive simulation studies. We apply our approach to Penn Medicine EHRs to identify patients under-diagnosed with non-alcoholic steatohepatitis (NASH).

1:40 PM–2:00 PM Speaker: Xin Zhou, Yale University

Doubly robust outcome-weighted learning for optimal treatment regimes

Author(s): Xin Zhou, Yale University; Michael Kosorok, University of North Carolina at Chapel Hill

Precision medicine is of considerable interest to clinical, academic and regulatory parties. The key to precision medicine is the optimal treatment regime. In this work, we propose the doubly robust outcome-weighted learning (DROL), which is semiparametric efficient and also has the desirable property of double robustness, to estimate the optimal treatment regime. Double robustness estimates the causal effect of a treatment on an outcome by combining both the outcome model and a propensity score model, which protects from misspecification in either model. We show that DROL is universally consistent, that is, the estimated regime of DROL converges the Bayes regime when the sample size approaches infinity, without knowing any specifics of the distribution of the data. We also propose variable selection methods for linear and nonlinear regimes, respectively, to further improve performance. The performance of the proposed DROL methods is illustrated in simulation studies and in an analysis of electronic health records data for the individualized treatment recommendation of the use of transthoracic echocardiography for intensive care unit patients with sepsis in MIMIC-III.

2:00 PM–2:20 PM Speaker: Shuang Wang, Columbia University

Integrating genetic data and electronic health records for personalized health risk predictions using a graph convolution network

Author(s): Yuqi Miao, Columbia University; Chunhua Weng, Columbia University; Krzysztof Kiryluk, Columbia University; Shuang Wang, Columbia University

Electronic Health Records (EHR) and molecular profiles such as genetic or gene expression data of patients are complementary types of health information that is valuable for personalized health risk predictions. However, due to their vastly different data dimensions, it is methodologically challenging to integrate them efficiently and effectively for prediction purposes. Here we propose GCN-EPI: a Graph Convolution Network for EHR data and Patient graph from molecular profiles Integration. With patient graphs from patient molecular profiles and through graph convolution, GCN-EPI aggregates EHR data and molecular profiles by weighting EHR information using neighboring patients through patient graphs adaptively to improve personalized health risk predictions. Our simulation studies suggested superior prediction performance of GCN-EPI over that of competing methods including Lasso and random forest on concatenated features from molecular profiles and EHR, and similarity network fusion (SNF) that fuses two similarity matrices from molecular profiles and EHR. We applied the proposed GCN-EPI algorithm and competing methods to predict incident end stage kidney disease (ESKD) using genetic data and EHR data extracted from the data warehouse of Columbia University Irving Medical Center and observed improved prediction performance of the proposed GCN-EPI.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S17: RECENT ADVANCES IN INTEGRATIVE STATISTICAL GENETICS AND GENOMICS

Monday, June 17, 2024
1:00 PM–2:30 PM, Blackbird B (Mezzanine Level)

Organizer: Hongkai Ji, Johns Hopkins University
Chair: Kevin Lin, University of Washington

1:00 PM–1:20 PM Speaker: Xiang Zhou, University of Michigan

Accurate and efficient integrative reference-informed spatial domain detection for spatial transcriptomics

Author(s): Ying Ma, Xiang Zhou

Spatially resolved transcriptomics (SRT) studies are becoming increasingly common and large, offering unprecedently opportunity in mapping complex tissue structures and functions. Here, we present IRIS, a computational method designed to characterize tissue spatial organization in SRT studies through accurately and efficiently detecting spatial domains. IRIS uniquely leverages single-cell RNA-seq data for reference-informed detection of biologically interpretable spatial domains, integrating multiple SRT slices while explicitly considering correlations both within and across slices. We demonstrate the advantages of IRIS through in-depth analysis of six SRT datasets encompassing diverse technologies, tissues, species, and resolutions. In these applications, IRIS achieves significant accuracy gains (39%-1,083%) and speed improvements (4.6-666.0) in moderate-sized datasets, while representing the only method applicable for large datasets including stereo-seq and 10x Xenium. As a result, IRIS reveals intricate brain structures, uncovers tumor microenvironment heterogeneity, and detects structural changes in diabetes-affected testis, all at a speed and accuracy unachievable by existing approaches.

1:20 PM–1:40 PM Speaker: Zhana Duren, Clemson University

Inferring gene regulatory networks from single cell multiome data using atlas-scale external data

Author(s): Qiuyue Yuan, Clemson University; Zhana Duren, Clemson University

Existing methods for Gene Regulatory Networks (GRNs) inference rely on gene expression data alone, or on lower resolution bulk data. Despite recent integration of ATAC-seq and RNA-seq data, learning complex mechanisms from limited independent data points still presents a daunting challenge. Here we present LINGER (LIfelong neural Network for GEne Regulation), a machine learning method to infer GRNs from single-cell paired gene expression and chromatin accessibility data. LINGER incorporates both atlas-scale external bulk data across diverse cellular contexts and prior knowledge of transcription factor (TF) motifs as a manifold regularization. LINGER achieves 4-7-fold relative increase in accuracy over existing methods and reveals a complex regulatory landscape of genome-wide association studies, enabling enhanced interpretation of disease-associated variants and genes. Following the GRN inference from a reference sc-multiome data, LINGER allows for the estimation of TF activity solely from bulk or single-cell gene expression data, leveraging the abundance of available gene expression data to identify driver regulators from case-control studies.

1:40 PM–2:00 PM Speaker: Kelly Street, University of Southern California

Multi-omic analysis methods for identifying phenotypic plasticity

Author(s): Kelly Street, University of Southern California; Yifan Zhang, University of Southern California; Kimberly Siegmund, University of Southern California; Darryl Shibata, University of Southern California

Plasticity is a cell's ability to rapidly and reversibly alter its phenotype, typically without significant epigenetic remodeling. While plasticity is known to play an important role in many human systems, such as colon crypt maintenance, it may also help to explain how a single progenitor cell can give rise to a tumor with many different cellular phenotypes. We found evidence of phenotypic plasticity by examining DNA methylation profiles and single-cell RNA-Seq data from both normal colon crypts and colorectal cancers. Specifically, we found that genes with higher expression variability tended to have more conserved, or less variable, methylation profiles. However, many existing quantitative methods for the analysis of multi-omic data are designed to identify correlation or covariance between epigenetic features and gene expression. Such methods may not be well-calibrated to detect phenotypic plasticity, in which conservation, not variability, of the epigenome plays an important role.

2:00 PM–2:20 PM Speaker: Mengjie Chen, University of Chicago

Beyond variability: a novel gene expression stability metric to unveil homeostasis and regulation

Author(s): Mengjie Chen, University of Chicago

A homeostatic cell carries out regular functions to maintain balance within the cell. This involves continuously reacting to both internal and external stimuli, a process that often involves the regulation of transcription. Within a homeostatic cell, the majority of genes are transcriptionally stable, and a smaller proportion might be in a regulatory or compensatory state, acting as a response to stimuli. Key candidates for this type of regulation could be 'first responder' genes, interferons and genes that encode heat shock proteins, among many others. Here, we introduce the concept of gene expression stability, measured by the gene homeostasis Z-index. This index unveils genes subject to precise regulation within specific cell subsets, shedding light on their roles in cellular adaptation. For example, we observe organ-specific patterns exemplified by heightened synaptic transmission activities in islets. Furthermore, we uncover regulatory patterns for neuropeptides, such as insulin and somatostatin, exhibiting extreme values within a limited number of cells. These findings underscore the limitations of conventional mean-based approaches, highlighting our approach's ability to surpass these constraints.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S18: IMPACT OF SPATIOTEMPORAL MODELING ON REAL-WORLD DYNAMICS

Monday, June 17, 2024
1:00 PM–2:30 PM, Gold (Lower Level)

Organizer: Indranil Sahoo, Virginia Commonwealth University
Chair: Indranil Sahoo, Virginia Commonwealth University

1:00 PM–1:20 PM Speaker: Meng Li, Rice University

Just plug in: Optimal Gaussian process modeling for rate of change

Author(s): Zejian Liu, Rice University; Meng Li, Rice University

The rate of change is often an interesting nonparametric function in many applications. In the Bayesian paradigm, Gaussian processes (GPs) are routinely used as a flexible prior for unknown functions, and are arguably one of the most popular tools in many areas including spatio-temporal data analysis. However, little is known about the optimal modeling strategy and theoretical properties when using GPs for derivatives (rate of change), and misconceptions exist. We study a plug-in strategy by differentiating the posterior distribution with GP priors for derivatives of any order. We obtain posterior contraction rates for plug-in GPs and demonstrate their remarkable adaptability to derivative orders. We also establish minimax optimality and provide a data-driven hyperparameter tuning method. This leads to a practically simple nonparametric Bayesian method with optimal and adaptive hyperparameter tuning for simultaneously estimating the regression function and its derivatives. Climate change applications will be discussed.

1:20 PM–1:40 PM Speaker: Suman Majumder, University of Missouri

Multivariate cluster point process to quantify and explore multi-entity configurations: Application to biofilm image data

Author(s): Suman Majumder, University of Missouri; Brent A. Coull, Harvard T.H. Chan School of Public Health; Jessica L. Mark Welch, Forsyth Institute; Patrick J. La Riviere, University of Chicago; Floyd E. Dewhirst, Forsyth Institute; Jacqueline R. Starr, Brigham and Women's Hospital; Kyu Ha Lee, Harvard T.H. Chan School of Public Health

Clusters of similar or dissimilar objects are encountered in many fields. Frequently used approaches treat the central object of each cluster as latent. Yet, often objects of one or more types cluster around objects of another type. Such arrangements are common in biomedical images of cells, in which nearby cell types likely interact. Quantifying spatial relationships may elucidate biological mechanisms. Parent-offspring statistical frameworks can be usefully applied even when central objects (parents) differ from peripheral ones (offspring). We propose the novel multivariate cluster point process (MCPP) to quantify multi-object (e.g., multi-cellular) arrangements. Unlike commonly used approaches, the MCPP exploits locations of the central parent object in clusters. It accounts for possibly multilayered, multivariate clustering. The model formulation requires specification of which object types function as cluster centers and which reside peripherally. If such information is unknown, the relative roles of object types may be explored by comparing fit of different models via the deviance information criterion (DIC). In simulated data, we compared DIC of a series of models; the MCPP correctly iden-

tified simulated relationships. It also produced more accurate and precise parameter estimates than the classical univariate Neyman-Scott process model. We also used the MCPP to quantify proposed configurations and explore new ones in human dental plaque biofilm image data. MCPP models quantified simultaneous clustering of Streptococcus and Porphyromonas around Corynebacterium and of Pasteurellaceae around Streptococcus and successfully captured hypothesized structures for all taxa. Further exploration suggested the presence of clustering between Fusobacterium and Leptotrichia, a previously unreported relationship.

1:40 PM–2:00 PM Speaker: Indranil Sahoo, Virginia Commonwealth University

Estimating atmospheric motion winds from satellite image data using space-time drift models

Author(s): Indranil Sahoo, Virginia Commonwealth University; Joe Guinness, Cornell University; Brian J. Reich, North Carolina State University

State-of-the-art AI applications and statistical challenges: Here, we explore the impact of AI, particularly deep learning, on medical imaging, and address the accompanying statistical challenges, such as data quality and model interpretability.

2:00 PM–2:20 PM Speaker: Ting Fung Ma, University of South Carolina

Hierarchical dependence modeling for the analysis of large insurance claims data

Author(s): Ting Fung Ma, University of South Carolina; Yizhou Cai, University of South Carolina; Peng Shi, University of Wisconsin-Madison; Jun Zhu, University of Wisconsin-Madison

Extreme weather events associated with climate change have caused significant damages. In particular, hail storms damage millions of properties in the U.S. resulting in billion-dollar insured losses each year in the recent decade. To facilitate the insurance claims management operations in insurance companies, we construct a hierarchical dependence model which accommodates the complex dependence within and between the outcomes of interests including the propensity of filing a claim, time to report a claim, and the claim amount. The storm-specific and property-specific characteristics are incorporated through marginal models, such as generalized linear models and survival analysis models. The dependence within the hail event is captured by spatial factor copula, while the dependence between different outcomes is captured by bivariate copula. For parameter estimation, we develop a two-step procedure that first maximizes the marginal likelihood function and then maximizes the pairwise likelihood, which ensures computational feasibility for big data. We apply this modeling framework to analyze a large dataset involving hail storms in Colorado from 2011 to 2015 impacting hundreds of thousands of insured properties and demonstrate that the predictive performance can be improved by our proposed methodology.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S19: UNLEASHING THE POWER OF STATISTICS IN COMPANION DIAGNOSTICS (CDX) AND DRUG CO-DEVELOPMENT PROCESS

Monday, June 17, 2024
1:00 PM–2:30 PM, Melody (Lobby Level)

Organizer: Hong Wang, Sanofi
Chair: Xiaoli Kong, Wayne State University

1:00 PM–1:20 PM Speaker: Siena Tabuena-Frolli, Agilent Technologies

Study design and statistical considerations to meet differing requirements from CMDE and FDA

Author(s): Siena Tabuena-Frolli, Agilent Technologies & California State University, Long Beach; Malinka Jansson, Agilent Technologies; Karina Kulangara, Agilent Technologies; Grace Lee, Agilent Technologies

Companion diagnostic (CDx) devices are key to furthering the success of precision medicine, necessitating global regulatory approvals for novel therapies and corresponding CDx devices. Different regulatory bodies, such as the US Federal Drug Administration (FDA) and China's National Medical Products administration (NMPA), have varying requirements for demonstrating medical device safety and effectiveness. While some requirements for demonstrating precision and robustness of devices overlap, others, such as accepted study designs and analysis methods, differ. This discussion will focus on the similarities and differences in CDx device submission requirements for the US and China markets, from a statistical standpoint

1:20 PM–1:40 PM Speaker: Shuguang Huang, Stat4ward

Sensitivity analysis in CDx Bridging Study

Author(s): Hong Wang, Sanofi; Shuguang Huang, Stat4ward; Siena Tabuena-Frolli, Agilent

Enrichment design is commonly used in oncology clinical trials where patients were enrolled if the clinical trial assay (CTA) indicated the biomarker status is positive (CTA+). When the trial is shown to be successful, a bridging study is required to bridge the CTA to a market-ready assay (MRA) that is called companion diagnostic assay (CDx). The goal of the bridging study is to demonstrate that the clinical efficacy would have been maintained if the CDx had been used in the trial. Using statistical annotations with Y representing clinical benefit and E representing the expected value, the goal of the bridging study is to demonstrate: $E(Y|CTA+) = E(Y|CDx+)$

Since CTA negative (CTA-) patients were not enrolled in the trial, the clinical efficacy for (CTA-CDx+) patients is missing. A typical approach is to do a sensitivity analysis by assuming that $E(Y_CTA - CDx+) = c_E(Y_CTA + CDx+)$ where *c* value is to vary between 0 and 1.

This talk will discuss the relationship between the *c* value (unknown) and the concordance (observed), typically measured by PPA and NPA, between CTA and CDx.

1:40 PM–2:00 PM Speaker: Hong Wang, Sanofi

Clinical decision marking rule guided cut-point selection for companion diagnostic (CDx) development

Author(s): Hong Wang, Shen Zhang, Shuguang Huang, Yiding Zhang, Wenting Wang

Determining an appropriate cut-point for continuous biomarker is critical in drug and companion diagnostic (CDx) co-development process. The traditional methods aim to select cut point to maximize the clinical efficacy in selected subgroup. However, these methods tend to select extreme cutpoint with low biomarker prevalence and result in (1) difficulty in patient enrollment during trial conduct, and (2) potential patients may miss appropriate treatment opportunity.

To address this challenge, we proposed to use clinical decision marking rule to guide cutpoint selection for CDx development. We used real clinical trial to demonstrate the concept and the implement process. In the simulation study, we implemented the three-outcome decision making (3ODM) framework based continuous assessment during the Proof of Concept (PoC) stage for simultaneous cutpoint selection and efficacy evaluation. This novel 3ODM guided approach aims preserve sufficient ability to achieve the Proof of Concept (PoC) criteria determined by 3ODM framework and maximize target population to efficiently bring the right treatment to the right patients who can benefit most from it. The clinical decision marking rule guided cut-point selection strategy allows pre-match CDx development timeline with clinical trial timeline to benefit smooth and accelerate development from different perspective including regulatory, trial operation, assay development, statistical considerations, et al.

2:00 PM–2:20 PM Speaker: Jie Cheng, Takeda Pharmaceutical Company

Novel approaches for patient subgroup identification and predictive gene signature development

Author(s): Jie Cheng, Takeda Pharmaceutical Company; Jacob Zhang, Takeda Pharmaceutical Company; Robin Mogg, Takeda Pharmaceutical Company; Richard Labotka, Takeda Pharmaceutical Company

We will present our machine learning based approaches for two common tasks in clinical stage drug development: exploratory patient subgroup identification using patient baseline variables (demographics, clinical biomarkers etc.), and predictive gene signature development from high dimensional array data. For both tasks, the goal is to identify patient subgroups where the treatment effect is significantly larger than the rest of the study population. We will use a failed phase 3 oncology study to showcase the utility of our machine learning based approaches.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S20: Best Student Paper Awards: Honorable Mentions

Monday, June 17, 2024
1:00 PM–2:30 PM, Green Room (Lobby Level)

Organizer: Ran Tao, Vanderbilt University Medical Center
Chair: Siyuan Ma, Vanderbilt University Medical Center

1:00 PM–1:22 PM Speaker: Sydney Louit, University of Connecticut

CALF-SBM: A covariate-assisted latent factor stochastic block model

Author(s): Sydney Louit, University of Connecticut; Jun Yan, University of Connecticut; Panpan Zhang, Vanderbilt University; Evan Clark, Vanderbilt University; Niketna Vivek, Vanderbilt University; Alexander Gelbard, Vanderbilt University

We propose a novel network generative model extended from the standard stochastic block model by concurrently utilizing observed node-level information and accounting for network-enabled nodal heterogeneity. The proposed model is so called covariate-assisted latent factor stochastic block model (CALF-SBM). The inference for the proposed model is done in a fully Bayesian framework. The primary application of CALF-SBM in the present research is focused on community detection, where a model-selection-based approach is employed to estimate the number of communities which is practically assumed unknown. To assess the performance of CALF-SBM, an extensive simulation study is carried out, including comparisons with multiple classical and modern network clustering algorithms. Lastly, the paper presents two real data applications, respectively based on an extremely new network data demonstrating collaborative relationships of otolaryngologists in the United States and a traditional aviation network data containing information about direct flights between airports in the United States and Canada.

1:22 PM–1:44 PM Speaker: Yangjianchen Xu, University of North Carolina at Chapel Hill

Checking the Cox proportional hazards model with interval-censored data

Author(s): Yangjianchen Xu, University of North Carolina at Chapel Hill; Donglin Zeng, University of Michigan; Danyu Lin, University of North Carolina at Chapel Hill

This paper presents a general framework for checking the adequacy of the Cox proportional hazards model with interval-censored data. Specifically, we construct certain stochastic processes that are informative about various aspects of the model, i.e., proportional hazards assumption, functional forms of covariates and exponential link function. We show that these stochastic processes can be viewed as the score statistics for testing zero regression parameters under some extended models. We establish their weak convergence to Gaussian processes through modern empirical process theory. We then approximate the limiting distributions by Monte Carlo simulation and develop graphical and numerical procedures to check model assumptions and improve goodness-of-fit. We evaluate the performance of the proposed methods through extensive simulation studies and provide an application to the Atherosclerosis Risk in Communities Study.

1:44 PM–2:06 PM Speaker: Travis Canida, University of Maryland

Multivariate Bayesian variable selection with application to multi-trait genetic fine mapping

Author(s): Travis Canida, University of Maryland; Hongjie Ke, University of Maryland; Shuo Chen, University of Maryland; Zhenyao Ye, University of Maryland; Tianzhou Ma, University of Maryalnd

Variable selection has played a critical role in modern statistical learning and scientific discoveries. Numerous regularization and Bayesian variable selection methods have been developed in the past two decades for variable selection, but most of these methods consider selecting variables for only one response. As more data is being collected nowadays, it is common to analyze multiple related responses from the same study. Existing multivariate variable selection methods select variables for all responses without considering the possible heterogeneity across different responses, i.e. some features may only predict a subset of responses but not the rest. Motivated by the multi-trait fine mapping problem in genetics to identify the causal variants for multiple related traits, we developed a novel multivariate Bayesian variable selection method to select critical predictors from a large number of grouped predictors that target at multiple correlated and possibly heterogeneous responses. Our new method is featured by its selection at multiple levels, its incorporation of prior biological knowledge to guide selection and identification of best subset of responses predictors target at. We showed the advantage of our method via extensive simulations and a real fine mapping example to identify causal variants associated with different subsets of addictive behaviors.

2:06 PM–2:28 PM Speaker: Shiyu (Richard) Shu, The George Washington University

Longitudinal benefit:risk analysis through the desirability of outcome ranking (DOOR) with application to ACTT-1 Trial

Author(s): Shiyu Shu, The George Washington University; Guoqing Diao, The George Washington University; Toshimitsu Hamasaki, The George Washington University; Scott Evans, The George Washington University

A clinical trial aims to uncover high-quality evidence for both efficacy and safety for an experimental treatment. Traditionally, statisticians evaluate the two aspects marginally, where, for example, efficacy considers the proportion of patients cured and safety considers the proportion of patients not showing adverse events. However, analyzing the two aspects separately could end up with misleading results, as the cumulative nature of clinical outcomes and the correlation between efficacy and safety endpoints are often neglected. The desirability of outcome ranking (DOOR) method addresses such issues and provides a patient-centric approach to benefit:risk evaluation. A patient's outcome is ranked based on pre-specified clinical criteria, where the most desirable rank represents a good outcome with no side effects and the least desirable rank is the worst possible clinical outcome. The DOOR probability, estimated with a type of U-statistic, is used to evaluate the treatment effect. As the DOOR outcome is a temporal state that can have measures at multiple time points, we propose a longitudinal approach that estimates and infers the temporal treatment effects. We develop a methodology for constructing simultaneous confidence bands by accounting for the correlations across different time points. Additionally, we propose a weighted Mann-Whitney-U statistic to evaluate the treatment effect over the entire trial period. The performance of the proposed methodologies is examined through simulations and an application to a COVID-19 trial.

P2: EMPOWERING PATHS: MENTORSHIP AND CAREER DEVELOPMENT FOR WOMEN

Monday, June 17, 2024
3:00 PM–4:30 PM, Symphony 1 (Lobby Level)

Organized and moderated by Qingxia "Cindy" Chen, Vice Chair of Education at Vanderbilt University Medical Center's Department of Biostatistics, this panel of leading women statisticians will share their observations and insights on professional development and building connections. The panelists are:

- Yu Shen, Chair ad interim of Biostatistics, MD Anderson Cancer Center
- Grace Y. Yi, Canada Research Chair in Data Science, University of Western Ontario
- Xun Chen, Vice President, Global Head of Biostatistics & Programming, Sanofi
- Margaret Gamalo Vice President, Statistics Therapeutic Area Head, Pfizer
- Jennifer Clark, Lead Mathematical Statistician, FDA

S21: CHALLENGES AND EXPLORATIONS OF ADVANCED STATISTICAL METHODOLOGIES IN THE REALM OF SURVIVAL ANALYSIS

Monday, June 17, 2024
3:00 PM–4:30 PM, Blackbird A (Mezzanine Level)

Organizer: Wenqing He, University of Western Ontario
Chair: Peijun Sang, University of Waterloo

3:00 PM–3:20 PM Speaker: David Oakes, University of Rochester Medical Center

Extending the proportional odds model to multiple event-time data

Author(s): David Oakes, University of Rochester Medical Center

It is well known (c.f. Cox and Oakes, 1984, p. 79) that a family of univariate survival distributions indexed by a continuous parameter satisfies both the accelerated life model and the proportional odds model if and only if the distributions follow the log-logistic form. Post hoc analysis of the bladder tumor data quoted by Wei, Lin and Weissfeld (1989) revealed a surprisingly good fit for a proportional odds model with the number of tumors as an ordered categorical outcome variable and the logarithm of follow-up time entered as an additional explanatory variable with unit coefficient (offset). Motivated by this example we will investigate formulations extending the proportional odds model to data on multiple events.

3:20 PM–3:40 PM Speaker: Liqun Diao, University of Waterloo

Poly trees for survival data

Author(s): Liqun Diao, University of Waterloo; Yixing Zhao, University of Waterloo

Pólya trees are commonly used as priors in nonparametric Bayesian analysis. This presentation will discuss approaches for utilizing Pólya trees to characterize the distribution of time-to-event data, which may be subjected to different forms of censoring such as right censoring or interval censoring. The discussion will cover different aspects of Pólya trees, including partitions, prior strength, and choices of prior distributions. Comparisons of the proposed methods to existing approaches for estimating survival probabilities are provided in both simulated settings and through applications to real datasets. It is shown that the proposed methods either improve upon or remain competitive with existing nonparametric estimation methods.

3:40 PM–4:00 PM Speaker: Yun-Hee Choi, Western University

Trivariate joint modeling for family data with longitudinal counts, recurrent events and a terminal event with application to Lynch Syndrome

Author(s): Jingwei Lu, Western University; Grace Y. Yi, Western University; Denis Rustand, King Abdullah University of Science and Technology; Patrick Parfrey, Memorial University of Newfoundland; Laurent Briollais, Lunenfeld-Tanenbaum Research Institute; University of Toronto; Yun-Hee Choi, Western University

The application of trivariate joint modeling for longitudinal count data, recurrent events, and a terminal event for family data has increased interest in medical studies. For example, families with Lynch Syndrome (LS) are at high risk of developing colorectal cancer (CRC), where

the number of polyps and the frequency of colonoscopy screening visits are highly associated with the risk of CRC among individuals and families. To assess how screening visits influence polyp detection, which itself influences time to CRC, we propose a clustered trivariate joint model. The proposed model is designed to accommodate longitudinal count data with zero-inflation and overdispersion and dependence among subjects and families via subject-specific and family-specific random effects. Our model is formulated as a latent Gaussian model, which allows us to use the Bayesian estimation approach with the Integrated Nested Laplace Approximation algorithm. The trivariate joint model was applied to a series of 18 families from Newfoundland, Canada. The occurrence of CRC was taken as a terminal event, colonoscopy screening visits as recurrent events, and the number of polyps detected at each visit as zero-inflated count data with overdispersion. Our findings demonstrated that the trivariate model fitted better than alternative bivariate models and that cluster effects should not be ignored when analyzing family data. Finally, the proposed model enables us to quantify heterogeneity across families and individuals in polyp detection and CRC risk. This, in turn, allows us to identify individuals and families that would benefit from more intensive screening visits.

4:00 PM–4:20 PM Speaker: Li-Pang Chen, National Chengchi University

Boosting method for length-biased and interval-censored survival data subject to high-dimensional error-prone covariates

Author(s):Li-Pang Chen, National Chengchi University

In this talk, we consider the length-biased and partly interval-censored data, whose challenges primarily come from biased sampling and interfere induced by interval censoring. Unlike existing methods that focus on low-dimensional data and assume the covariates to be precisely measured, sometimes researchers may encounter high-dimensional data subject to measurement error, which are ubiquitous in applications and make estimation unreliable. To address those challenges, we explore a valid inference method for handling high-dimensional length-biased and interval-censored survival data with measurement error in covariates under the accelerated failure time model. We primarily employ the SIMEX method to correct for measurement error effects and propose the boosting procedure to do variable selection and estimation. The proposed method is able to handle the case that the dimension of covariates is larger than the sample size and enjoys appealing features that the distributions of the covariates are left unspecified

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S22: ADVANCING CLINICAL TRIAL ANALYSIS: INTEGRATING EXTERNAL DATA AND NOVEL ESTIMATION TECHNIQUES

Monday, June 17, 2024
3:00 PM–4:30 PM, Lyric (Lobby Level)

Organizer: Shu Yang, North Carolina State University
Chair: Xiaofei Wang, Duke University

3:00 PM–3:20 PM Speaker: Bo Zhang, Fred Hutchinson Cancer Center

Generalizing the intention-to-treat effect of an active control against placebo from historical placebo-controlled trials to an active-controlled trial: A case study of the efficacy of daily oral TDF/FTC in the HPTN 084 study

Author(s): Qijia He, University of Washington; Fei Gao, Fred Hutchinson Cancer Center; Oliver Dukes, Ghent University; Sinead Delany-Moretlwe, University of the Witwatersrand; Bo Zhang, Fred Hutchinson Cancer Center

In many clinical settings, an active-controlled trial design (e.g., a non-inferiority or superiority design) is often used to compare an experimental medicine to an active control (e.g., an FDA-approved, standard therapy). One prominent example is a recent phase 3 efficacy trial, HIV Prevention Trials Network Study 084 (HPTN 084), comparing long-acting cabotegravir, a new HIV pre-exposure prophylaxis (PrEP) agent, to the FDA-approved daily oral tenofovir disoproxil fumarate plus emtricitabine (TDF/FTC) in a population of heterosexual women in 7 African countries. One key complication of interpreting study results in an active-controlled trial like HPTN 084 is that the placebo arm is not present and the efficacy of the active control (and hence the experimental drug) compared to the placebo can only be inferred by leveraging other data sources. In this article, we study statistical inference for the intention-to-treat (ITT) effect of the active control using relevant historical placebo-controlled trials data under the potential outcomes (PO) framework. We highlight the role of adherence and unmeasured confounding, discuss in detail identification assumptions and two modes of inference (point versus partial identification), propose estimators under identification assumptions permitting point identification, and lay out sensitivity analyses needed to relax identification assumptions. We applied our framework to estimating the intention-to-treat effect of daily oral TDF/FTC versus placebo in HPTN 084 using data from an earlier Phase 3, placebo-controlled trial of daily oral TDF/FTC (Partners PrEP).

3:20 PM–3:40 PM Speaker: Amir Asiaee, Vanderbilt University Medical Center

Leveraging observational data for efficient CATE estimation in randomized controlled trials

Author(s): Amir Asiaee, Vanderbilt University Medical Center; Chiara Di Gravio, Imperial College; Yuting Mei, Vanderbilt University; Jared D. Huling, University of Minnesota

Randomized controlled trials (RCTs) are the gold standard for causal inference, but they are often powered only for average effects, making estimation of heterogeneous treatment effects (HTEs)challenging. Conversely, large-scale observational studies (OS) offer a wealth of data but suffer from confounding bias. Our paper presents a novel framework to leverage OS data for enhancing the efficiency in estimating conditional average treatment effects (CATEs) from RCTs while mitigating common biases. We propose an innovative approach to combine RCTs

and OS data, expanding the traditionally used control arms from external sources. The framework relaxes the typical assumption of CATE invariance across populations, acknowledging the often unaccounted systematic differences between RCT and OS participants. We demonstrate this through the special case of a linear outcome model, where the CATE is sparsely different between the two populations. The core of our framework relies on learning potential outcome means from OS data and using them as a nuisance parameter in CATE estimation from RCT data. We further illustrate through experiments that using OS findings reduces the variance of the estimated CATE from RCTs and can decrease the required sample size for detecting HTEs.

3:40 PM–4:00 PM Speaker: Jiwei Zhao, University of Wisconsin-Madison

Doubly safe estimation for the average treatment effect on the treated with external control data under high-dimensionality

Author(s): Jiwei Zhao, University of Wisconsin-Madison

Randomized controlled trial (RCT) has been a gold standard for causal discovery in various biomedical studies. In this paper, we consider the situation that some external control data, possibly with a much larger sample size, are available. However, the standard doubly robust estimator for ATT incorporating external controls might be even less efficient than the naïve doubly robust estimator without using the external controls. This is not ideal because it means the incorporation of external controls might be harmful for our estimation. To fix this issue, we propose a novel doubly robust estimator which is guaranteed to be always more efficient than the naïve doubly robust estimator without using the external controls. Further, if all models are correct, the proposed estimator is the same as the standard doubly robust estimator incorporating external controls, and it is semiparametrically efficient. The asymptotic theory developed in this paper, including both estimation and statistical inference, is under the general high-dimensional confounder situation. We conduct comprehensive simulation studies, as well as a real data application, to illustrate our proposed methodology.

4:00 PM–4:20 PM Speaker: Peisong Han, Gilead Sciences, Inc.

Improving prediction of linear regression models by integrating external information from heterogeneous populations: James-Stein estimators

Author(s):Peisong Han, Haoyue Li, Sung Kyun Park, Bhramar Mukherjee, Jeremy Taylor

We consider the setting where (i) an internal study builds a linear regression model for prediction based on individual-level data, (ii) some external studies have fitted similar linear regression models that use only subsets of the covariates and provide coefficient estimates for the reduced models without individual-level data, and (iii) there is heterogeneity across these study populations. The goal is to integrate the external model summary information into fitting the internal model to improve prediction accuracy. We adapt the James-Stein shrinkage method to propose estimators that have guaranteed improvement in the prediction mean squared error after information integration, regardless of the degree of study population heterogeneity.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S23: STATISTICAL AND CAUSAL INFERENCE WITH IRREGULARLY SPACED OBSERVATION TIMES

Monday, June 17, 2024
3:00 PM–4:30 PM, Ocean Way (Mezzanine Level)

Organizer: Shu Yang, North Carolina State University
Chair: Shu Yang, North Carolina State University

3:00 PM–3:20 PM Speaker: Yujing Gao, North Carolina State University

Semi-parametric sensitivity analysis for trials with irregular and informative assessment times

Author(s): Bonnie B. Smith, Johns Hopkins Bloomberg School of Public Health; Yujing Gao, North Carolina State University; Shu Yang, North Carolina State University; Ravi Varadhan, Johns Hopkins School of Medicine; Andrea J. Apter, University of Pennsylvania Perelman School of Medicine; Daniel O. Scharfstein,University of Utah School of Medicine

Many trials are designed to collect outcomes at or around pre-specified times after randomization. In practice, there can be substantial variability in the times at which participants are actually assessed. Such irregular assessment times pose a challenge to learning the effect of treatment since not all participants have outcome assessments at the times of interest. Furthermore, in some trials, the observed outcome values may not be representative of all participants' outcomes at a given time. This problem, known as informative assessment times, can arise if participants tend to have assessments at times when their outcomes are better (or worse) than at other times, or if participants with better outcomes tend to have more (or fewer) assessments. Methods have been developed that account for some types of informative assessment; however, since these methods rely on untestable assumptions, sensitivity analyses are needed. Here, we develop a sensitivity analysis methodology by extending existing weighting methods. Our method accounts for the possibility that participants with worse outcomes at a given time are more (or less) likely than other participants to have an assessment at that time, even after controlling for variables observed earlier in the study. For example, in an asthma trial, participants may be less (or more) likely to attend a data collection appointment when they have a severe asthma exacerbation. We apply our method to a randomized trial of low-income individuals with uncontrolled asthma. The implementation of our influence-function based estimation procedure is illustrated in detail. We also derive the large-sample distribution of our estimator and evaluate the finite-sample performance of our sensitivity analysis procedure in a realistic simulation study.

3:20 PM–3:40 PM Speaker: Wei Jin, Johns Hopkins University

A Bayesian decision framework for optimizing sequential combination antiretroviral therapy in people with HIV

Author(s): Wei Jin, Johns Hopkins University; Yang Ni, Texas A&M University; Jane O'Halloran, Washington University in St. Louis; Amanda Spence, Georgetown University; Leah Rubin, Johns Hopkins University School of Medicine; Yanxun Xu, Johns Hopkins University

Numerous adverse effects (e.g., depression) have been reported for combination antiretroviral therapy (cART) despite its remarkable success in viral suppression in people with HIV. To improve long-term health outcomes for people with HIV, there is an urgent need to design per-

sonalized optimal cART with the lowest risk of comorbidity in the emerging field of precision medicine for HIV. Large-scale HIV studies offer researchers unprecedented opportunities to optimize personalized cART in a data-driven manner. However, the large number of possible drug combinations for cART makes the estimation of cART effects a high-dimensional combinatorial problem, imposing challenges in both statistical inference and decision-making. We develop a two-step Bayesian decision framework for optimizing sequential cART assignments. In the first step, we propose a dynamic model for individuals' longitudinal observations using a multivariate Gaussian process. In the second step, we build a probabilistic generative model for cART assignments and design an uncertainty-penalized policy optimization using the uncertainty quantification from the first step. Applying the proposed method to a dataset from the Women's Interagency HIV Study, we demonstrate its clinical utility in assisting physicians to make effective treatment decisions, serving the purpose of both viral suppression and comorbidity risk reduction.

3:40 PM–4:00 PM Speaker: Janie Coulombe, Université de Montréal

Comparison of imputation and inverse-weighting approaches for the estimation of causal effects in longitudinal data

Author(s): Janie Coulombe, Université de Montréal; Erica E. M. Moodie, McGill University; Susan M. Shortreed, Kaiser Permanente Washington

We have access to electronic health records data from Kaiser Permanente Washington (KPW) in the US. With these data, we created a cohort of patients who initiated antidepressant drugs between 2008 and 2018 and who have a confirming diagnosis of depression. We aim to estimate the average causal effect of citalopram vs fluoxetine, two commonly prescribed selective serotonin reuptake inhibitors, on the severity of depression. That severity is measured via the patient health questionnaire which is recorded irregularly in time and across patients. Its observation is assumed to be associated with patient characteristics. We describe two approaches to estimate consistently the average causal effect in that setting: one based on multiple imputations, and one based on inverse intensity of visit weighting. The approaches are both compared in large simulation studies and in the application to the KPW dataset. We also discuss their advantages and pitfalls.

4:00 PM–4:20 PM Speaker: Luo Xiao, North Carolina State University

Functional data analysis for longitudinal data with informative observation times

Author(s): Caleb Weaver, North Carolina State University; Luo Xiao, North Carolina State University; Wenbin Lu, North Carolina State University

In functional data analysis for longitudinal data, the observation process is typically assumed to be noninformative, which is often violated in real applications. Thus, methods that fail to account for the dependence between observation times and longitudinal outcomes may result in biased estimation. For longitudinal data with informative observation times, we find that under a general class of shared random effect models, a commonly used functional data method may lead to inconsistent model estimation while another functional data method results in consistent and even rate-optimal estimation. Indeed, we show that the mean function can be estimated appropriately via penalized splines and that the covariance function can be estimated appropri-

ately via penalized tensor-product splines, both with specific choices of parameters. For the proposed method, theoretical results are provided, and simulation studies and a real data analysis are conducted to demonstrate its performance.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S24: CHALLENGES AND ADVANCES IN RISK ASSESSMENT AND PREDICTION

Monday, June 17, 2024
3:00 PM–4:30 PM, Platinum (Lower Level)

Organizer: Jing Ning, University of Texas MD Anderson Cancer Center
Chair: Jian Wang, University of Texas MD Anderson Cancer Center

3:00 PM–3:20 PM Speaker: Chixiang Chen, University of Maryland, School of Medicine

Improving estimation efficiency for survival data analysis by integrating a coarsened time-to-event outcome from an external study

Author(s): Chixiang Chen, Daxuan Deng, Ming Wang

In the era of big data, increasing availability of data makes combining different data sources to obtain more accurate estimations a popular topic. However, the development of data integration is often hindered by the heterogeneity in data forms across studies. In this paper, we focus on a case in survival analysis where we have primary study data with a continuous time-to-event outcome and complete covariate measurements, while the data from an external study contain an outcome observed at regular intervals and only a subset of covariates is measured. To incorporate external information while accounting for the different data forms, we posit working models and obtain informative weights by empirical likelihood, which will be used to construct a weighted estimator in the main analysis. We have established the theory demonstrating that the new estimator has higher estimation efficiency compared to the conventional ones, and this advantage is robust to working model misspecification, as confirmed in our simulation studies. To assess its utility, we apply our method to accommodate data from the National Alzheimer's Coordinating Center to improve the analysis of the Alzheimer's Disease Neuroimaging Initiative phase 1 study.

3:20 PM–3:40 PM Speaker: Frank Harrell, Vanderbilt University Medical Center

Ordinal state transition models as a unifying risk prediction framework

Author(s): Frank Harrell, Vanderbilt University Medical Center

In this talk I will present a case for the use of discrete time Markov ordinal longitudinal state transition models as a unifying approach to modeling a variety of outcomes for the purpose of estimating risk and expected time in a given state, and for comparing treatments in clinical trials. This model structure can be used to analyze time until a single terminating event, longitudinal binary events, recurrent events, continuous longitudinal data, and longitudinal ordinal responses including multiple events. Partial information can be formally incorporated using standard likelihood approaches without the need for imputation. The model also provides a formal way to assess evidence for consistency of a treatment effect over different outcomes.

3:40 PM–4:00 PM Speaker: Jin Piao, University of Southern California

Semiparametric isotonic regression modelling and estimation for group testing data

Author(s): Ao Yuan, Georgetown University; Jin Piao, University of Southern California; Jing Ning, University of Texas MD Anderson Cancer Center; Jing Qin, National Institute of Allergy and Infectious Diseases

In the group testing procedure, several individual samples are grouped and the pooled samples, instead of each individual sample, are tested for outcome status (e.g., infectious disease status). Although this cost-effectiveness strategy in data collection is both labor and time efficient, it poses statistical challenges to derive statistically and computationally efficient estimators under semiparametric models. We consider semiparametric isotonic regression models for the simultaneous estimation of the conditional probability curve and covariate effects, in which a parametric form for combining the covariate information is assumed and the monotonic link function is left unspecified. We develop an expectation-maximization algorithm to overcome the computational challenge and embed the pool-adjacent violators algorithm in the M-step to facilitate the computation. We establish the large sample behavior of the proposed estimators and examine their finite sample performance in simulation studies. We apply the proposed method to data from the National Health and Nutrition Examination Survey for illustration.

4:00 PM–4:20 PM Speaker: Baojiang Chen, University of Texas Health Science Center at Houston

Maximum profile binomial likelihood estimation for the semiparametric Box-Cox power transformation model

Author(s): Pengfei Li, University of Waterloo; Tao Yu, National University of Singapore; Baojiang Chen, University of Texas Health Science Center at Houston; Jing Qin, National Institutes of Health

The Box-Cox transformation model has been widely applied for many years. The parametric version of this model assumes that the random error follows a parametric distribution, say the normal distribution, and estimates the model parameters using the maximum likelihood method. The semiparametric version assumes that the distribution of the random error is completely unknown; existing methods either need strong assumptions or are less effective when the distribution of the random error significantly deviates from the normal distribution. We adopt the semiparametric assumption and propose a maximum profile binomial likelihood method. We theoretically establish the joint distribution of the estimators of the model parameters. Through extensive numerical studies, we demonstrate that our method has an advantage over existing methods when the distribution of the random error deviates from the normal distribution. Furthermore, we compare the performance of our method and existing methods on an HIV data set.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S25: RECENT ADVANCES IN DESIGN AND ANALYSIS OF CLINICAL TRIALS WITH TIME-TO-EVENT OUTCOMES

Monday, June 17, 2024
3:00 PM–4:30 PM, Sound Emporium A/B (Mezzanine Level)

Organizer: Gaohong Dong, BeiGene Chair: Huan Cheng, BeiGene

3:00 PM–3:20 PM Speaker: Milind Phadnis, University of Kansas Medical Center

Sample size calculation for two-arm trials with survival endpoint using the concept of Relative Time for the non-proportional hazards scenario

Author(s): Milind Phadnis, University of Kansas Medical Center

Sample size calculations for two-arm clinical trials with a time-to-event endpoint have traditionally used the assumption of proportional hazards (PH) or the assumption of exponentially distributed survival times. Available software provides methods for sample size calculation using a nonparametric logrank test, Schoenfeld's formula for Cox PH model, or parametric calculations specific to the exponential distribution. In cases where the PH assumption is not valid, the first-choice method is to compute sample size assuming a piecewise linear survival curve (Lakatos approach) for both the control and treatment arms with judiciously chosen cut-points. Some newer methods have been developed in the context of AFT models thereby allowing non-proportional hazards. These methods, however, always assume an instantaneous effect of treatment relative to control requiring that the effect size be defined by a single number whose magnitude is preserved throughout the trial duration. Here, we consider the scenarios where the hypothesized benefit of treatment relative to control may not be constant giving rise to the notion of Relative Time (RT). By assuming that survival times for control and treatment arm come from two different Weibull distributions with different location and shape parameters, we develop the methodology for sample size calculation for specific cases of both non-PH and non-AFT. We also demonstrate an application for a real-world example. Further, we show how this method can be extended to allow group sequential testing as well as stochastic curtailment tests.

3:20 PM–3:40 PM Speaker: Qing Zhao, Merck & Co., Inc.

Calculation of phase 2 dose-finding study sample size for reliable phase 3 dose selection

Author(s): Fang Liu, Merck; Qing Zhao, Merck; Anthony Rodger, Merck; Devan Mehtrotra, Merck

Sample sizes of phase 2 dose-finding studies, usually determined based on a power requirement to detect a significant dose-response relationship, will generally not provide adequate precision for phase 3 target dose selection. We propose to calculate the sample size of a dose finding study based on the probability of successfully identifying the target dose within an acceptable range (e.g., 80%-120% of the target) using the multiple comparison and modeling procedure (MCP-Mod). With the proposed approach, different design options for the phase 2 dose finding study can also be compared. Due to inherent uncertainty around an assumed true dose-response relationship, sensitivity analyses to assess robustness of the sample size calculations to deviations from modeling assumptions are recommended. Planning for a hypothetical phase 2 dose finding study is used to illustrate the main points. Codes for the proposed approach are available at https://github.com/happysundae/posMCPMod.

3:40 PM–4:00 PM Speaker: Yu Cheng, University of Pittsburgh

Interim analysis in sequential multiple assignment randomized trials for survival outcomes

Author(s): Zi Wang, University of Pittsburgh; Yu Cheng, University of Pittsburgh; Abdus Wahed, University of Pittsburgh

Sequential Multiple Assignment Randomized Trials (SMARTs) have been conducted to mimic the actual treatment processes experienced by physicians and patients in clinical settings and inform comparative effectiveness of dynamic treatment regimes (DTRs). In a SMART design, patients are involved in multiple stages of treatment, and the treatment assignment is adapted over time based on the patient's characteristics such as disease status and treatment history. In this work, we develop and evaluate statistically valid interim monitoring (IM) approaches to allow for early termination of SMART trials for efficacy regarding time-to-event outcomes. The development is nontrivial. First, in comparing estimated event rates from different DTRs, log-rank statistics need to be carefully weighted to account for overlapping treatment paths. At a given time point, we can then test for the null hypothesis of no difference among all DTRs based on a weighted log-rank type statistic. With multiple stages, the number of DTRs is much larger than the number of treatments involved in a typical randomized trial, resulting in many parameters to estimate for the covariance matrix of the weighted log-rank statistics. More challengingly, for IM, we need to quantify how the log-rank statistics at two different time points are correlated, and each component of the covariance matrix depends on a mixture of event processes which can jump at multiple time points due to the nature of multiple assignments. Efficacy boundaries at multiple interim analyses can then be established using the Pocock and the O'Brien Fleming (OBF) boundaries. We run extensive simulations to evaluate and compare type I error and power for our proposed weighted log-rank Chi-square statistic for DTRs under different boundary specifications. The methods are demonstrated in analyzing a neuroblastoma trial.

4:00 PM–4:30 PM **Q&A and Floor Discussion**

S26: KISS SESSION I: RECENT ADVANCES IN THE DESIGN AND ANALYSIS OF CLINICAL TRIALS

Monday, June 17, 2024
3:00 PM–4:30 PM, Southern Ground A/B (Mezzanine Level)

Organizer: Yeonhee Park, University of Wisconsin-Madison
Chair: Leena Choi, Vanderbilt University Medical Center
KISS = Korean International Statistical Society

3:00 PM–3:20 PM Speaker: Mi-Ok Kim, University of California, San Francisco

Bayesian adaptive design for covariate-adaptive historical control information borrowing

Author(s): Huaqing Jin, University of California San Francisco; Mi-Ok Kim, University of California San Francisco; Aaron Scheffler, University of California San Francisco; Fei Jiang, University of California San Francisco

Interest in incorporating historical data in the clinical trial has increased with the rising cost of conducting clinical trials. The intervention arm for the current trial often requires prospective data to assess a novel treatment, and thus borrowing historical control data commensurate in distribution to current control data is motivated in order to increase the allocation ratio to the current intervention arm. Existing historical control borrowing adaptive designs adjust allocation ratios based on the commensurability assessed through study-level summary statistics of the response agnostic of the distributions of the trial subject characteristics in the current and historical trials. This can lead to distributional imbalance of the current trial subject characteristics across the treatment arms as well as between current control data and borrowed historical control data. Such covariate imbalance may threaten the internal validity of the current trial by introducing confounding factors that affect study endpoints. In this article, we propose a Bayesian design which borrows and updates the treatment allocation ratios both covariate-adaptively and commensurate to covariate dependently assessed similarity between the current and historical control data. We employ covariate-dependent discrepancy parameters which are allowed to grow with the sample size and propose a regularized local regression procedure for the estimation of the parameters. The proposed design also permits the current and the historical controls to be similar to varying degree, depending on the subject level characteristics. We evaluate the proposed design extensively under the settings derived from two placebo-controlled randomized trials on vertebral fracture risk in post-menopausal women.

3:20 PM–3:40 PM Speaker: Deukwoo Kwon, University of Texas Health Science Center at Houston

Power calculation for detecting interaction effect in cross-sectional stepped-wedge cluster randomized trials: An important tool for disparity research

Author(s): Chen Yang, Icahn School of Medicine at Mount Sinai; Asem Berkalieva, Icahn School of Medicine at Mount Sinai; Madhu Mazumdar, Icahn School of Medicine at Mount Sinai; Deukwoo Kwon, University of Texas Health Science Center at Houston

Stepped-Wedge Cluster-Randomized Trials (SW-CRTs) are increasingly utilized for evaluating complex healthcare delivery interventions where simple CRTs are not feasible. Appealing features of SW-CRTs include having each cluster acting as their own control, not needing to withhold the intervention from any patient, and having time to prepare clusters for administra-

tion of intervention while collecting baseline information. However, the design and analysis of SW-CRT is complex and methodology is not available for many scenarios including detection of interaction effects. Detecting interaction effect is important for a variety of research scenarios. We present several ways of computing power and showcase their comparative performance through simulation. We then apply the methodology to a published SW-CRT with binary outcome.

3:40 PM–4:00 PM Speaker: Kyungbok Lee, Seoul National University

Contextual bandit algorithm with multiple stochastically correlated outcomes

Author(s): Kyungbok Lee, Seoul National University; Myunghee Cho Paik, Shepherd23 Inc.; Min-hwan Oh, Seoul National University; Gi-Soo Kim, UNIST

We present a novel variant of a contextual bandit problem with multi-dimensional reward feedback formulated as a mixed-effects model, where the correlations between multiple feedback are induced by sharing stochastic coefficients called random effects. We propose a novel algorithm, Mixed-Effects Contextual UCB (ME-CUCB), achieving $\tilde{O}(d\sqrt{mT})$ regret bound after T rounds where d is the dimension of contexts and m is the dimension of outcomes, with either known or unknown covariance structure. This is a tighter regret bound than that of the naive canonical linear bandit algorithm ignoring the correlations among rewards. We prove a lower bound of $\Omega(d\sqrt{mT})$ matching the upper bound up to logarithmic factors. To our knowledge, this is the first work providing a regret analysis for mixed-effects models and algorithms involving weighted least-squares estimators. We provide numerical experiments demonstrating the advantage of our proposed algorithm, supporting the theoretical claims.

4:00 PM–4:20 PM Speaker: Yeonhee Park, University of Wisconsin-Madison

Data-driven monitoring for phase II clinical trial designs based on percentile event time test

Author(s): Yeonhee Park, University of Wisconsin-Madison

The goal of phase II clinical trials is to evaluate the therapeutic efficacy of a new drug. Some investigators want to use the time-to-event endpoint as the primary endpoint of the phase II study to see the improvement of the therapeutic efficacy of a new drug in median survival time. Recently, median event time test (METT) has been proposed to provide a simple and straightforward rule which compares the observed median survival time with the prespecified threshold. However, median survival time would not be observed during the trial if the drug performs well and indeed cures most patients or if the accrual rate is so fast. To address the issues in clinical practice, we first propose a percentile event time test (PETT), which generalizes METT to any percentile of the survival time, and develop data-driven monitoring for phase II clinical trial designs based on PETT. We illustrate the proposed method with a trial example andevaluate the performance of the method through simulations.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S27: STATISTICAL MARVELS: UNVEILING INSIGHTS FROM HIGH-DIMENSIONAL COMPLEX DATA IN HEALTHCARE PRACTICE

Monday, June 17, 2024
3:00 PM–4:30 PM, Blackbird B (Mezzanine Level)

Organizer: Wenyu Gao, University of North Carolina at Charlotte
Chair: Qingning Zhou, University of North Carolina at Charlotte

3:00 PM–3:20 PM Speaker: Huang Lin, University of Maryland

MetVAE: A novel deep learning framework for confounding correction and molecular co-occurrence inference in metabolomics data

Author(s): Huang Lin, University of Maryland; James Morton, Gutz Analytics

Metabolomics, bridging genomic output and environmental input, is a vital component in the "omics" cascade, offering unique insights into disease diagnosis, personalized treatment strategies, and environmental health studies. However, similar to other omics data types, metabolomics data are characterized by high-dimensionality and compositional nature. Moreover, large-scale metabolomics studies often involve confounding factors such as batch processing, which introduces significant biases within the generated data. To navigate these challenges, we introduce the Metabolomics Variational Autoencoder (MetVAE), a state-of-the-art extension of the deep learning algorithm, Variational Autoencoder (VAE). MetVAE model is specially tailored to address the intricacies of metabolomics data, by both adhering to the compositional nature of untargeted metabolomics data and handling technical bias, notably the confounding factors. MetVAE offers an integrated solution for metabolomics data analysis, enabling dimension reduction and providing robust inferences on molecular co-occurrence patterns of metabolites with the presence of confounders. Comprehensive simulations and real data applications validate MetVAE' efficiency and accuracy, paving the way for more robust, precise, and comprehensive analysis in metabolomics research.

3:20 PM–3:40 PM Speaker: Han Chen, Virginia Tech

Joint variables selection for finite mixture of functional regression models

Author(s): Han Chen, Virginia Tech; Inyoung Kim, Virginia Tech

In this study, we address the challenge of variable selection within a finite mixture of regression models (FMR) and a finite mixture of functional regression models (FMFR) when dealing with high-dimensional, heterogeneous data. Recognizing that regression models in FMR/FMFR often share a common set of predictors, our approach introduces a joint variable selection method. This method leverages a shared structure, employing a penalized log-likelihood framework to borrow strength across the regression components. Additionally, we extend this methodology to mixtures of functional regressions and develop penalized expectation-maximization (EM) algorithms for effective numerical optimization. Our primary goal is to enhance the efficiency of variable selection in addressing unknown heterogeneous population issues, drawing inspiration from autism spectrum disorder (ASD) studies. The proposed method shows potential in identifying distinct subpopulations and elucidating their unique characteristics. We demonstrate the advantages of our approach through rigorous numerical analyses, which include both simulated data and an empirical study focusing on brain imaging in ASD.

3:40 PM–4:00 PM Speaker: Yabo Niu, University of Houston

Covariate-assisted Bayesian graph learning for heterogeneous data

Author(s): Yabo Niu, University of Houston; Yang Ni, Texas A&M University; Debdeep Pati, Texas A&M University; Bani K. Mallick, Texas A&M University

In traditional Gaussian graphical models, data homogeneity is routinely assumed with no extra variables affecting the conditional independence. In modern genomic datasets, there is an abundance of auxiliary information, which often gets under-utilized in determining the joint dependency structure. In this talk, I will present a new Bayesian approach to model undirected graphs underlying heterogeneous multivariate observations with additional assistance from covariates. Building on product partition models, we propose a novel covariate-dependent Gaussian graphical model that allows graphs to vary with covariates so that observations whose covariates are similar share a similar undirected graph. To efficiently embed Gaussian graphical models into the proposed framework, we explore both Gaussian likelihood and pseudo-likelihood functions. For Gaussian likelihood, a G-Wishart prior is used as a natural conjugate prior, and for the pseudo-likelihood, a product of Gaussian-conditionals is used. Moreover, the proposed model induced by the prior has large support and is flexible to approximate any piece-wise constant conditional variance-covariance matrices. Furthermore, based on the theory of fractional likelihood, the rate of posterior contraction is minimax optimal assuming the true density to be a Gaussian mixture with a known number of components. I will demonstrate the efficacy of the approach via numerical studies and analysis of protein networks for a breast cancer dataset assisted by genetic covariates.

4:00 PM–4:20 PM Speaker: Wenyu Gao, University of North Carolina at Charlotte

Informed weighted Dirichlet process mixture for functional clustering in highly correlated high-dimensional data

Author(s): Wenyu Gao, University of North Carolina at Charlotte; Inyoung Kim, Virginia Tech

Functional clustering in high-dimensional data poses challenges, especially in scenarios with unknown cluster counts. While nonparametric Bayesian methods such as the Dirichlet process mixture (DPM) model offer approaches, they often do not effectively leverage observational information. Conversely, the weighted Dirichlet process mixture (WDPM) model incorporates prior information via a weight function. However, its investigation remains limited, particularly in functional clustering. This study explores informed weight functions for WDPM in functional clustering, addressing the gap in research by exploring covariates beyond Euclidean distances. We apply this method to fMRI data from autism spectrum disorder (ASD) patients, integrating spatial correlations and demographic information to enhance clustering accuracy.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S28: METHODS AND APPLICATIONS OF NETWORK DATA SCIENCE: A SHOWCASE OF THE JOURNAL OF DATA SCIENCE

Monday, June 17, 2024
3:00 PM–4:30 PM, Gold (Lower Level)

Organizer: Jun Yan, University of Connecticut
Chair: Panpan Zhang, Vanderbilt University Medical Center

3:00 PM–3:20 PM Speaker: Yuguo Chen, University of Illinois Urbana-Champaign

Modularity based methods for network data

Author(s): Yuguo Chen, University of Illinois Urbana-Champaign

We introduce several network modularity measures for both single-layer and multi-layer networks under different null models of the network, motivated by empirical observations in networks from a diverse field of applications. We describe a statistical framework for modularity-based network community detection. The effectiveness of the proposed methods is demonstrated in simulated and real networks.

3:20 PM–3:40 PM Speaker: Yunpeng Zhao, Colorado State University

Variational estimators of the degree-corrected latent block model

Author(s): Yunpeng Zhao, Colorado State University; Ning Hao, University of Arizona; Ji Zhu, University of Michigan, Ann Arbor

Biclustering on bipartite graphs is an unsupervised learning method that simultaneously groups two types of objects in the graph, such as users and movies in a movie review dataset. The latent block model (LBM) has been proposed as a model-based tool for biclustering. However, the effectiveness of the LBM is often limited by the influence of row and column sums in the data matrix. To address this limitation, we introduce the degree-corrected latent block model (DC-LBM), which accounts for the varying degrees in row and column clusters, significantly enhancing performance on real-world datasets and simulated data. We develop an efficient variational expectation-maximization algorithm by creating closed-form solutions for parameter estimates in the M steps. Furthermore, we prove the label consistency and the rate of convergence of the variational estimator under the DC-LBM, allowing the expected graph density to approach.

3:40 PM–4:00 PM Speaker: Yuan Zhang, The Ohio State University

U-statistic reduction: Higher-order accurate risk control and statistical-computational trade-off, with application to network method-of-moments

Author(s): Meijia Shao, Meta; Dong Xia, Hong Kong University of Science and Technology; Yuan Zhang, The Ohio State University

U-statistics play central roles in many statistical learning tools but face the haunting issue of scalability. Significant efforts have been devoted into accelerating computation by U-statistic reduction. However, existing results almost exclusively focus on power analysis, while little work addresses risk control accuracy—comparatively, the latter requires distinct and much

more challenging techniques. In this paper, we establish the first statistical inference procedure with provably higher-order accurate risk control for incomplete U-statistics. The sharpness of our new result enables us to reveal how risk control accuracy also trades off with speed for the first time in literature, which complements the well-known variance-speed trade-off. Our proposed general framework converts the long-standing challenge of formulating accurate statistical inference procedures for many different designs into a surprisingly routine task. This paper covers non-degenerate and degenerate U-statistics, and network moments. We conducted comprehensive numerical studies and observed results that validate our theory's sharpness. Our method also demonstrates effectiveness on real-world data applications.

4:00 PM–4:20 PM Speaker: Zhiyong Zhang, University of Notre Dame

Structural equation modeling with network data

Author(s): Zhiyong Zhang, University of Notre Dame; Ziqian Xu, University of Notre Dame

Social network analysis is useful for identifying the structural relationships among social entries. To promote the use of social network analysis and develop better methods for social and behavioral research, we propose a general structural equation modeling framework in which a social network can be treated as a new type of observed variables. Under the framework, we have investigated the roles of social networks as a predictor, an outcome, and a mediator. In this talk, I will present several studies to illustrate the application of the new framework.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S29: UNLOCKING THE POWER OF COVARIATES IN REGULATORY DECISION-MAKING: EXPLORING COVARIATE ADJUSTMENT, PROCOVA, DIGITAL TWINS, AND BEYOND

Monday, June 17, 2024
3:00 PM–4:30 PM, Melody (Lobby Level)

Organizer: Chenguang Wang, Regeneron Pharmaceuticals
Chair: Jeen (Jing-ou) Liu, Regeneron Pharmaceuticals

3:00 PM–3:20 PM Speaker: Rolando Acosta, Regeneron Pharmaceuticals

Developing digital biomarkers for auditory and vestibular phenotyping in clinical trials

Author(s): Rolando Acosta, Regeneron; Erin Robertson, Regeneron; Emily Redington, Regeneron; Meghan Drummond, Regeneron; Jacek Urbanek, Regeneron

Background: The development of therapeutics for hearing and balance has been a historical challenge. Commonly used outcome measures for hearing loss and vestibular dysfunction have several limitations, including time-consuming and costly diagnostic testing. Furthermore, subjective questionnaires may not reflect actual disability and may be confounded by neurocognitive impairment. Improved auditory and vestibular phenotyping is critical to enable large-scale natural history studies and the success of future clinical trials.

Objectives: (1) Data-driven identification of easy-to-measure risk factors related to hearing loss. (2) To create a digital composite covariate (DCC) to support the identification of patients who may potentially suffer from undiagnosed hearing loss.

Methods: We analyzed cross-sectional audiometry, body measures, hemodynamics, and questionnaire data for 22,000 subjects across nine cohorts from the National Health and Nutrition Examination Survey (NHANES). The outcome of interest was hearing loss and was captured in the form of the pure-tone average (PTA). Body measurements, hemodynamic values, and questionnaire information were used as predictors. We develop a stacking ensemble model between a LASSO and an XGBoost using the first seven NHANES cohorts. The stacking weights were computed using nonnegative linear regression. Model performance was evaluated using the root mean square error and Pearson correlation estimate between the predicted and observed PTA values. Model features were ranked based on their predictive ability. Model performance was evaluated in cohorts eight and nine.

Results: Our model yields a root mean square error of 7.21dB in the held-out cohort and a correlation of 0.76 between the predicted and observed PTA. Age, self-assessed hearing ability, and use of hearing aid were important predictors. Surprisingly, some body measurements, like waist circumference, were also important predictors of hearing loss. Lastly, sociodemographic features with important predictive abilities include biological sex, education level, and race.

3:20 PM–3:40 PM Speaker: Arman Sabbaghi, Unlearn.AI

Statistical methods for utilizing digital twins to deliver more efficient randomized controlled trials

Author(s): Arman Sabbaghi, Unlearn.AI

Randomized controlled trials (RCTs) are the gold standard for evaluating the causal effects of

new medical treatments, interventions, and therapies. However, modern trials are becoming increasingly difficult to conduct due to enrollment challenges, long trial durations, and significant costs. We present innovative statistical methodologies and novel trial designs from Unlearn.AI's Digital Twins Platform, which combine historical data, artificial intelligence (AI), and randomization to deliver smaller, faster, and more powerful RCTs that are built with regulatory guidance in mind. Central to the platform is a Digital Twin Generator (DTG), which is constructed by the application of generative AI algorithms on historical control data. The DTG takes the baseline data of participants from prospective RCTs as its inputs, and yields digital twin forecasts for the participants' future clinical outcomes under the control condition. Summaries from the digital twins are used to empower our new statistical methods for RCTs, while ensuring unbiased treatment effect estimators and Type I error rate control. Unlearn's Digital Twin Platform and statistical methods can ultimately power the next generation of RCTs and accelerate effective decision-making for drug development, as they incorporate the unique advantages of DTGs to address the challenges of modern trials while maintaining alignment with regulatory guidance.

3:40 PM–4:00 PM Speaker: Henry Wei, Regeneron Pharmaceuticals

Operational efficiency in clinical trial design

Author(s): Henry Wei, Regeneron Pharmaceuticals

Clinical trial design incorporates many elements, including biomedical scientific, biostatistical, and operational considerations. Covariate adjustment methods offer an important way to improve the biostatistical efficiency of clinical trial design. These approaches are closely related to clinical trial enrichment methods, particularly disease-prognostic methods to address heterogeneity in the clinical trial populations. However, some biostatistical approaches may seem efficient, but result in major operational challenges. Refining a population of interest based on a single baseline characteristic or covariate, for example, may decrease the size of the recruitable population so much as to render the clinical trial delayed or otherwise no longer feasible. Similarly, accrual for multicenter clinical trials typically follows a sigmoidal curve. In industry-sponsored trials, due to the nature of clinical trial site activation and the presence for intermediary steps such as study site feasibility questionnaires, site contracting, site qualification visits, and site initiation visits. Therefore trade-off analyses to understand the impact of combinations of covariates, thresholds, and composite scores derived from multiple covariates may be helpful to identify combinations with maximum enrichment for minimum loss of recruitable population. Considering the multi-year lifecycle for drug development, it may be advantageous to conduct predictive modeling exercises in advance of Phase 2 or Phase 3 trial design needs to unearth novel covariates or composite covariates, enrichment strategies, and quantify trade-offs with operational efficiency.

4:00 PM–4:20 PM Discussant: Zhiwei Zhang, Gilead Sciences, Inc.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S30: TRANSFORMING THE DIGITAL TO THE REAL WORLD: EHR AND GRAPHICAL DATA APPROACHES TO ENHANCE PATIENT OUTCOMES

Monday, June 17, 2024
3:00 PM–4:30 PM, Green Room (Lobby Level)

Organizer: Erik Bloomquist, Merck & Co., Inc.
Chair: Heng Zhou, Merck & Co., Inc.

3:00 PM–3:20 PM Speaker: Zhijun Yin, Vanderbilt University Medical Center

The hidden patient connections: Predicting hormonal therapy medication discontinuation using hypergraph neural network on clinical communications

Author(s): Zhijun Yin, Vanderbilt University Medical Center

Hormonal therapy is an important adjuvant treatment for breast cancer patients, but medication discontinuation of such therapy is not uncommon. The goal of this work is to conduct research on the modeling of clinic communications, which have shown value in understanding medication discontinuation, to predict the discontinuation of hormonal therapy medications. Notably, we leveraged the Hypergraph Neural Network to capture the hidden connections of patients inferred from clinical communications. Combining the content of clinical communications as well as the demographics, insurance, and cancer stage information, our model achieved an AUC of 67.9%, which significantly outperformed other baselines such as Graph Convolutional Network (65.3%), Random Forest (62.7%), and Support Vector Machine (62.8%). Our study suggested that incorporating the hidden patient connections encoded in clinical communications into prediction models could boost their performance. Future research would consider combining structured medical records and clinical communications to better predict medication discontinuation.

3:20 PM–3:40 PM Speaker: Hulin Wu, University of Texas Health Science Center at Houston

EHR data analysis and prediction: Fairness, stability and reliability

Author(s): Hulin Wu, University of Texas Health Science Center at Houston

Use of real world EHR data for research is promising, but challenging due to some typical features of EHR data: sampling bias, uncertainty in diagnosis and data missing not at random. Thus, the prediction and analysis results can be biased, unstable and unreliable if these EHR data problems cannot be appropriately addressed. In this talk, I will illustrate the problems and propose the methods for handling these problems using application examples from HIV and diabetes studies based on a nationwide EHR database.

3:40 PM–4:00 PM Speaker: Chuan Hong, Duke University

Evaluating generative large language models in healthcare

Author(s): Chuan Hong, Duke University; Monica Agarwal, Duke University; Armando Bedoya, Duke University; Anand Chowdhury, Duke University; Anthony Sorrentino, Duke University; Sophia Bessias, Duke University; Nicoleta Economou-Zavlanos, Duke University; Fan Li, Duke University; Eric Poon, Duke University; Michael Pencina, Duke University

The rapid evolution of large language models (LLMs) has ushered in a new era of computational linguistics, yet a systematic approach to their evaluation, particularly in sensitive domains such as healthcare, remains nascent. This work bridges these gaps by offering a detailed and integrated review of qualitative evaluation, quantitative evaluation, and meta-evaluation. For quantitative evaluation, our review introduces a taxonomy of evaluation metrics, categorizing them based on essential dimensions such as human supervision, contextual data, and analytical depth. In addition to generic settings, our work distinctively emphasizes additional considerations vital in the healthcare sector. As a result, we propose an integrated cross-walk between qualitative and quantitative assessment methods. The proposed framework harmonizes qualitative insights, such as user-focused evaluations, with objective quantitative metrics. We present a detailed "go-to menu" of evaluation criteria, tailored to address specific healthcare applications and emphasize distinct aspects in both pre-deployment and post-deployment phases. Our findings underscore the need for evaluations that extend beyond mere technical accuracy, factoring in medical ethics, fairness, equity, and potential operational biases. Our work offers a summary of existing methods of LLM evaluation that can establish a baseline from which future evaluation methods can be developed to keep pace with the rapid advancements in the field.

4:00 PM–4:20 PM Speaker: Xu Shi, University of Michigan

Harmonizing electronic health record and claims data across FDA Sentinel Initiative Data Partners: Case study and lessons learned

Author(s): Xu Shi, University of Michigan

The US Food and Drug Administration (FDA) Sentinel Initiative is a national surveillance system with a distributed data network of electronic health records (EHR) and claims data on >100 million patient lives from 17 data partners to monitor the safety of FDA-regulated medical products. The Sentinel System uses the Sentinel Common Data Model to standardize data elements and unify the medical coding "vocabulary" across participating sites. However, the coding "dialect" (i.e., the use and interpretation of codes) may still differ due to heterogeneity in care practice and financial drivers. With increasingly diverse data partners and medical coding systems, there is more and more variation in the way a clinical concept can be coded. Existing manually curated medical code ontology and mapping are not scalable and are error-prone. Data sharing constraints bring additional challenges. In this talk, we present data-driven and privacy-preserving statistical methods for detecting and reducing coding differences between healthcare systems. We share our findings from a case study of data harmonization between two Sentinel data partners among a diabetic population.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

Poster Session and Mixer

Monday, June 17, 2024
6:00 PM–8:00 PM, Starstruck Gallery (Mezzanine Level)
Chair and Organizer: Fei Ye, Vanderbilt University Medical Center

View posters and enjoy refreshments while socializing with other conference participants. Poster awards will be presented at the Tuesday evening banquet.

Poster Competition Entries

Presenter: Kun Bai, Vanderbilt University Medical Center
Prognostic value of systemic inflammatory biomarkers detections in patients undergoing CRS-HIPEC by machine learning methods
Author(s): Kun Bai, Vanderbilt University Medical Center; Fei Ye, Vanderbilt University Medical Center; David Hanna, Vanderbilt University Medical Center; Deepa Magge, Vanderbilt University Medical Center

Laboratory biomarkers have been utilized as prognostic markers in various solid tumors. The objective of this study is to assess the potential of three laboratory values: blood neutrophil to lymphocyte ratio (NLR), platelet to lymphocyte ratio (PLR), and monocyte to lymphocyte ratio (MLR), as prognostic biomarkers in patients with peritoneal carcinomatosis who undergo cytoreductive surgery with hyperthermic intraperitoneal chemotherapy (CRS-HIPEC). We conducted a retrospective analysis of 156 patients who underwent CRS-HIPEC between 2013–2020. To predict survival (OS), recurrence-free survival (RFS), and postoperative outcomes, we used XGboost to train our machine-learning model and select the biomarkers. The accuracy of biomarker detection was compared between XGboost and traditional methods.

Presenter: Zhe Chen, University of Illinois Urbana-Champaign
Enhanced inference for distributions of individual treatment effects in randomized experiments and quasi-experiments
Author(s): Zhe Chen, University of Illinois Urbana-Champaign; Xinran Li, University of Chicago

Randomization-based inference has been a cornerstone for drawing valid causal inference from experimental data. Recent works have extended beyond conventional Fisher's and Neyman's approaches regarding constant or average treatment effects, and proposed valid statistical tests for inferring distributions of individual treatment effects. In this paper, we propose two novel methods that further enhance the existing inference for individual treatment effect distributions in completely and stratified randomized experiments, which could be overly conservative since it considers the unlikely worst-case scenario where units with large individual effects are all allocated to the treatment group. Specifically, our first method considers inference for individual effect distributions in treated and control groups, separately, and combine them to infer treatment effects for all experimental units. Our second method uses Berger and Boos's (1994) approach to control for the number of units with large effects that are assigned to treatment. Both simulation and applications demonstrate the substantial gain from these two improved methods. Finally, we extend the proposed improved methods to sensitivity analysis for quasi-experiments constructed through matching, and sampling-based randomized experiments where we want to generalize experimental evidence to larger populations of interest.

Presenter: Xin Gai, Duke University
Subtype-aware registration of longitudinal electronic health records
Author(s): Xin Gai, Duke University; Shiyi Jiang, Duke University; Anru Zhang, Duke University

A significant challenge in the analysis of electronic health record (EHR) data is the misalignment between the recorded start time of an event and the actual onset of the disease. Such misalignment can cause patients with similar disease progression having distinct trends in biomarkers, degrading the utility of downstream analyses, such as disease sub-phenotyping and risk prediction. In this study, we introduced an innovative method to enhance the alignment of electronic health record (EHR) data by integrating clustering and registration processes within a feedback mechanism. This approach focuses on iterative clustering and registration efforts to improve the accuracy of cluster centers and enhance registration precision, while simultaneously reducing computational and sample complexity. A distinctive feature of our method involves mapping longitudinal EHR data onto B-spline coefficient space, which facilitates the discrete optimization of clustering outcomes and the fine-tuning of curve alignments. Extensive simulations demonstrated that our method effectively restores true trajectories. Additionally, we validated our approach on various tasks using the MIMIC-IV dataset, finding that it significantly improves the performance of diverse downstream data analysis models.

Presenter: Xiaoming Gao, University of North Carolina at Chapel Hill
Multiple imputation of missing time-dependent covariates in interval-censored survival analysis
Author(s): Xiaoming Gao, University of North Carolina at Chapel Hill; Fei Zou, University of North Carolina at Chapel Hill; Qingxia Chen, Vanderbilt University School of Medicine

Multiple imputation is one of the most common approaches to analyzing incomplete data. Additional challenges arise when missingness is observed in covariates collected longitudinally, and the outcome of interest is interval-censored event time. A complete case analysis will result in reduced efficiency and could even generate biased parameter estimates. This paper proposes a time-sequential imputation method based on the idea of fully conditional specification to multiply impute missing time-dependent covariates in studies with interval-censored outcomes. In addition to covariates of interest, the imputation model also utilizes cumulative hazard estimated iteratively, risk status, and observance of failure in subjects' subsequent visits. Extensive simulation studies demonstrated improved performance over existing methods with reduced bias, improved efficiency, and valid inference. The proposed method was applied to the cohort data from the Atherosclerosis Risk in Communities (ARIC) study to assess the risk of hypertension with potential predictors.

Presenter: Hyeong Jin Hyun, Purdue University
Fast cost-constrained regression
Author(s): Hyeong Jin Hyun, Purdue University; Xiao Wang, Purdue University

The conventional statistical models assume the availability of covariates without associated costs, yet real-world scenarios often involve acquisition costs and budget constraints imposed on these variables. Scientists must navigate a trade-off between model accuracy and expenditure within these constraints. In this paper, we introduce fast cost-constrained regression (FCR), designed to tackle such problems with computational and statistical efficiency. Specifically, we develop fast and efficient algorithms to solve cost-constrained problems with the loss function

satisfying a quadratic majorization condition. We theoretically establish nonasymptotic error bounds for the algorithm's solution, considering both estimation and selection accuracy. We apply FCR to extensive numerical simulations and four datasets from the National Health and Nutrition Examination Survey. Our method outperforms the latest approaches in various performance measures, while requiring fewer iterations and a shorter runtime.

Presenter: Wanlin Juan, Medical College of Wisconsin

CCI: A consensus clustering-based imputation method for addressing dropout events in scRNA-seq data

Author(s): Wanlin Juan, Medical College of Wisconsin; Kwang Woo Ahn, Medical College of Wisconsin; Chien-Wei Lin, Medical College of Wisconsin

Single-cell RNA sequencing (scRNA-seq) is a cutting-edge technique in molecular biology and genomics, revealing the cellular heterogeneity. However, scRNA-seq data often suffers from dropout events, meaning that certain genes exhibit very low or even zero expression levels due to technical limitations. Existing imputation and denoising methods to address dropouts lack comprehensive evaluations in downstream analyses and do not show the robustness under different scenarios. In response to this challenge, we propose a consensus clustering-based imputation (CCI) method. CCI performs clustering on each subset sampled across genes and summarizes clustering outcomes to define cellular similarities. CCI leverages the information from similar cells and employs the similarities to impute gene expression levels. Our comprehensive evaluations demonstrate that CCI not only reconstructs the original data pattern, but also improves the performance of downstream analyses. CCI outperforms existing methods for data imputation under different scenarios, exhibiting remarkable robustness and generalization.

Presenter: Lingxuan Kong, University of Michigan

Adaptive risk-weighted learning for estimating optimal multiple-object dynamic treatment regime of kidney transplant

Author(s): Lingxuan Kong, University of Michigan; Kevin He, University of Michigan; Lu Wang, University of Michigan

Kidney transplant is widely recognized as a better treatment compared to dialysis for end-stage renal disease patients. However, the patients waiting for kidney transplants always outnumber the possible kidney donors, making the treatment unavailable all the time. Meanwhile, patients may be censored by death before they receive any kidney transplant. This special characteristic of kidney transplants makes it hard to provide any realistic and clear regime guiding physicians. Here we provide a novel general stage-wise solution for the allocation of limited treatments by proposing the adaptive risk-weighted learning method, which allocates treatments to the population who can benefit from the treatment and are at high risk of death at the end of each decision stage. A new search algorithm ensures the overall optimality of the estimated regime. We also generalized this method to multiple outcome scenarios such as viewing treatment survival and population treatment-free risk as two objects to make our method meet aggressive and conservative physicians' requirements by adjusting object weights.

Presenter: Wenrui Li, University of Pennsylvania

Graph-guided Bayesian factor model for integrative analysis of multi-modal data with noisy network information

Author(s): Wenrui Li, University of Pennsylvania; Qiyiwen Zhang, University of Pennsylvania; Qi Long, University of Pennsylvania

There is a growing body of literature on factor analysis that can capture individual and shared structures in multimodal data. However, few of these approaches incorporate biological knowledge such as functional genomics and functional metabolomics. Graph-guided statistical learning methods that can incorporate knowledge of underlying 100% networks have been shown to improve predication and classification accuracy, and yield more interpretable results. Moreover, these methods typically use graphs extracted from existing databases or rely on subject matter expertise which are known to be incomplete and may contain false edges. To address this gap, we propose a graph-guided Bayesian factor model that can account for network noise and identify globally shared, partially shared and modalityspecific latent factors in multi-modal data. Specifically, we use two sources of network information, including the noisy graph extracted from existing databases and the estimated graph from observed features in the dataset at hand, to inform the model for the true underlying network via a latent scale modeling framework. This model is coupled with the Bayesian factor analysis model with shrinkage priors to encourage feature-wise and modal-wise sparsity, thereby allowing feature selection and identification of factors of each type. We develop an efficient Markov chain Monte Carlo algorithm for posterior sampling. We demonstrate the advantages of our method over existing methods in simulations, and through analyses of gene expression and metabolomics datasets for Alzheimer's disease.

Presenter: Qinghua Lian, Medical College of Wisconsin

Multiple imputation of missing covariates in time-to-event data

Author(s): Qinghua Lian, Medical College of Wisconsin; Soyoung Kim, Medical College of Wisconsin; Michael Martens, Medical College of Wisconsin; Kwang Woo Ahn, Medical College of Wisconsin

Missing covariates frequently arise in clinical studies involving right-censored data. Multiple imputation (MI) is one of the most popular approaches for addressing this issue. In handling missing covariates for time-to-events data, adopting the censoring ignorable missingness at random (CIMAR) assumption is more appropriate than the commonly used missingness at random (MAR) assumption. CIMAR allows the missingness to be dependent only on the true event time and the observed covariates, but not on the censoring time. Existing MI methods have been proposed under MAR or missing-covariate-independent censoring assumption. However, this leads to biased results when the censoring times also depend on the covariates with missing values. Thus, we proposed the substantive model compatible fully conditional specification (SMC-FCS) MI method to time-to-event data assuming CIMAR and missing-covariate-dependent censoring. Our simulation study demonstrates that the proposed SMC-FCS MI method, combined with Rubin's rule, yields unbiased estimates and approximately correct 95% confidence interval coverage rates within both the framework of the Cox proportional hazard model and the proportional cause-specific hazards model.

Presenter: Michael Lightfoot, North Carolina State University

Pseudo-likelihood methods for censored functional data

Author(s): Michael Lightfoot, North Carolina State University; Madison Book, North Carolina State University; Ana-Maria Staicu, North Carolina State University

Analysis of sparse functional data has been primarily conducted under the assumption of an uninformative sampling design. Sparse functional principal component analysis (FPCA) produces consistent estimation in the mean and covariance functions as long as the missingness is random for each individual. When this assumption is violated, recent research pointed out bias in the model component estimation; see Weaver et al. (2022), although no such methods have investigated when the observed time sampling is dependent on the outcome variable. We examine a censoring process where the missingness is dependent on the longitudinal outcome. This case is particularly common in observational studies in biostatistics, as patients may drop out of studies due to circumstances related to their health. If the outcome variable of interest in the study is health-related, the missingness is therefore also dependent on the observational outcome. We find that when applying current methods for sparse functional data blindly to such data, the missingness introduces bias in the mean estimation, covariance estimation, and prediction of the missing observations. We propose a pseudo-likelihood formulation for the estimate of the mean function to adjust for the dependence of the missingness on the observational outcomes. We compare the performance of the proposed method with sparse FPCA in a simulation study and an application to real data.

Presenter: Jilei Lin, Georgetown University

Smoothed quantile regression for spatial data

Author(s): Jilei Lin, The George Washington University; Huixia Judy Wang, The George Washington University; Myungjin Kim, Kyungpook National University

Existing methods for spatial data either have difficulty capturing heterogeneous patterns over complex domains or overlook the heterogeneity in the tail of the response distribution, which often exhibits in social and health disparity studies, such as mortality rates and incomes. In this paper, we introduce a flexible quantile spatial model (QSM) framework, which can simultaneously capture spatial nonstationarity and heterogeneity via constant and spatially varying coefficients. This framework also allows researchers to study patterns across different tails of the response distribution depending on their research interests. We first present a QBiT estimation method based on bivariate penalized splines on triangulation. To further improve computational efficiency, we propose the SQBiT estimator by employing convolution smoothing in the loss function. The developed methods can effectively capture spatial nonstationarity, meanwhile preserving critical data features such as shape and smoothness across complex and irregular domains. Under some regularity conditions, we show that the proposed SQBiT estimator can achieve an optimal convergence rate under the L_2-norm. In addition, we establish the Bahadur representation of the estimator, which allows us to further establish the asymptotic normality for the constant coefficient estimator and construct asymptotic confidence intervals. For small samples, we propose an interval estimator for both constant and varying coefficients based on wild bootstrap. Through simulation studies, we demonstrate the numerical and computational advantages of SQBiT over existing methods. The application of SQBiT to study the spatial heterogeneity of US mortality demonstrates that the mortality rates depend on socioeconomic factors differently across space and the tails of the mortality distribution.

Presenter: Rufeng Liu, Florida State University

Bayesian density estimation on the product of simplexes and the hypercube using multivariate Bernstein polynomials

Author(s): Rufeng Liu, Andrés F. Barrientos, Claudia Wehrhahn, Alejandro Jara

We propose a Bayesian nonparametric model for density estimation on the product of simplex spaces and the hypercube. The model is particularly useful for cases where the available data consist of multiple compositional features alongside variables that take on values within bounded intervals. A compositional feature is a vector of non-negative components whose sum of values remains constant, such as the time an individual spends on different activities during the day or the fraction of different types of food consumed as part of a person's diet. Our approach relies on a generalization of random multivariate Bernstein polynomials and corresponds to a Dirichlet process mixture of products of Dirichlet and beta densities. Theoretical properties such as prior support and posterior consistency are studied. We evaluate the model's performance through a simulation study and a real-world application using data from the 2005–2006 cycle of the U.S. National Health and Nutrition Examination Survey (NHANES). Furthermore, the conditional densities derived under this modeling strategy can be used for regression analyses where both the response and predictors take values on the simplex space and/or hypercube.

Presenter: Yi Liu, North Carolina State University

Multi-source conformal inference under distribution shift

Author(s): Yi Liu, North Carolina State University; Alexander Levis, Carnegie Mellon University; Sharon-Lise Normand, Harvard University; Larry Han, Northeastern University

Recent years have experienced increasing utilization of complex machine learning models across multiple sources of data to inform more generalizable decision-making. However, distribution shifts across data sources and privacy concerns related to sharing individual-level data, coupled with a lack of uncertainty quantification from machine learning predictions, make it challenging to achieve valid inferences in multi-source environments. We consider the problem of obtaining distribution-free prediction intervals for a target population, leveraging multiple potentially biased data sources. We derive the efficient influence functions for the quantiles of unobserved outcomes in the target and source populations, and show that one can incorporate machine learning prediction algorithms in the estimation of nuisance functions while still achieving parametric rates of convergence to nominal coverage probabilities. Moreover, when conditional outcome invariance is violated, we propose a data-adaptive strategy to upweight informative data sources for efficiency gain and downweight non-informative data sources for bias reduction. We highlight the robustness and efficiency of our proposals for a variety of conformal scores and data-generating mechanisms via extensive synthetic experiments. Hospital length of stay prediction intervals for pediatric patients undergoing a high-risk cardiac surgical procedure between 2016-2022 in the U.S. illustrate the utility of our methodology.

Presenter: Fangzhi Luo, University of Georgia

Functional clustering for longitudinal associations between Social Determinants of Health and stroke mortality in the US

Author(s): Fangzhi Luo, University of Georgia; Jianbin Tan, Duke University; Donglan Zhang, NYU Grossman Long Island School of Medicine; Hui Huang, Renmin University of China; Ye Shen, University of Georgia

Understanding the longitudinally changing associations between Social Determinants of Health (SDOH) and stroke mortality is essential for effective stroke management. Previous studies have uncovered significant regional disparities in the relationships between SDOH and stroke mortality. However, existing studies have not utilized longitudinal associations to develop data-driven methods for regional division in stroke control. To fill this gap, we propose a novel clustering method to analyze SDOH – stroke mortality associations in US counties. To enhance the interpretability and statistical efficiency of the clustering outcomes, we introduce a new class of smoothness-sparsity pursued penalties for simultaneous clustering and variable selection in longitudinal associations. As a result, we can identify crucial SDOH that contribute to longitudinal changes in stroke mortality. This facilitates the clustering of US counties into different regions based on the relationships between these SDOH and stroke mortality. The effectiveness of our proposed method is demonstrated through extensive numerical studies. By applying our method to longitudinal data on SDOH and stroke mortality at the county level, we identify 18 important SDOH for stroke mortality and divide the US counties into two clusters based on these selected SDOH. Our findings unveil complex regional heterogeneity in the longitudinal associations between SDOH and stroke mortality, providing valuable insights into region-specific SDOH adjustments for mitigating stroke mortality.

Presenter: Pengfei Lyu, Florida State University

Replicability analysis of high dimensional data accounting for dependence

Author(s): Pengfei Lyu, Florida State University; Xianyang Zhang, Texas A&M University; Hongyuan Cao, Florida State University

Replicability is the cornerstone of scientific research. We study the replicability of data from high-throughput experiments, where tens of thousands of features are examined simultaneously. Existing replicability analysis methods either ignore the dependence among features or impose strong modeling assumptions, producing overly conservative or overly liberal results. Based on p-values from two studies, we use a four-state hidden Markov model to capture the structure of local dependence. Our method effectively borrows information from different features and studies while accounting for dependence among features and heterogeneity across studies. We show that the proposed method has better power than competing methods while controlling the false discovery rate, both empirically and theoretically. Analyzing datasets from genome-wide association studies reveals new biological insights that otherwise cannot be obtained by using existing methods.

Presenter: Danyang Skylar Shi, University of Washington

Copula mixture models for marked point processes in time

Author(s): Danyang Shi, University of Washington; Xiaotian Zheng, University of Wollongong

We propose a copula-based constructive framework for building marked point processes ob-

served over time. In the framework, a marked point process is characterized by the conditional multivariate density of the duration (interval between successive event times) and marks. We model the conditional multivariate density using a finite mixture of transition densities, each incorporating a different lagged duration and its associated marks. These mixture transition densities are constructed by exploiting the decomposition of a high-dimensional densities into bivariate-copula densities and marginal densities. Using bivariate copulas as building blocks allows for flexible modeling of multivariate non-Gaussian dependence among durations and marks. Our approach using copulas, combined with the mixture model formulation for the conditional multivariate density, yields a flexible model for high-order dependence in point processes with multivariate marks, while retaining computational efficiency. We investigate properties of the point process model analytically and through simulation studies. The methodology is illustrated with an example from the environmental science.

Presenter: Gefei Wang, Yale University

STitch3D: Construction of a 3D whole organism spatial atlas by joint modelling of multiple slices with deep neural networks

Author(s): Gefei Wang, Yale University and Hong Kong University of Science and Technology; Jia Zhao, Yale University and Hong Kong University of Science and Technology; Yan Yan, Hong Kong University of Science and Technology; Yang Wang, Hong Kong University of Science and Technology; Angela Wu, Hong Kong University of Science and Technology; Can Yang, Hong Kong University of Science and Technology

Spatial transcriptomics (ST) technologies are revolutionizing the way to explore the spatial architecture of tissues. Currently, ST data analysis is often restricted to a single two-dimensional (2D) tissue slice, limiting our capacity to understand biological processes that take place in 3D space. Here we present STitch3D, a unified framework that integrates multiple ST slices to reconstruct 3D cellular structures. By jointly modelling multiple slices and integrating them with single-cell RNA-sequencing data, STitch3D simultaneously identifies 3D spatial regions with coherent gene-expression levels and reveals 3D cell-type distributions. STitch3D distinguishes biological variation among slices from batch effects, and effectively borrows information across slices to assemble powerful 3D models. Through comprehensive experiments, we demonstrate STitch3D's performance in building comprehensive 3D architectures, which allow 3D analysis in the entire tissue region or even the whole organism. The outputs of STitch3D can be used for multiple downstream tasks, enabling a comprehensive understanding of biological systems. This method has been published in *Nature Machine Intelligence*.

Presenter: Zi Wang, University of Pittsburgh

Nonparametric estimation of subgroup mediation effects with semi-competing risks data

Author(s): Zi Wang, University of Pittsburgh; Yu Cheng, University of Pittsburgh

A treatment may have an effect on a non-terminal event (e.g., disease progression), which in turn may influence a terminal event (e.g., death), or the treatment may affect the terminal event directly. We thus are interested in evaluating the mediational effect of the treatment through the non-terminal event and the direct treatment effect on the terminal event. However, the conventional definition of natural direct effect and indirect effect is not appropriate here because of the semi-competing risks data structure, where time to a non-terminal event may be censored by a terminal event, but not vice versa. A principal stratification approach is adopted to define

the natural direct and indirect effects on one stratum and the total effect on all strata. We propose nonparametric estimators of the direct and indirect effects under suitable assumptions. The theoretical properties of the proposed estimators are established, and their good finite sample performance is illustrated through numerical studies.

Presenter: Debra Wetcher-Hendricks, Moravian University

An expanded McNemar approach for evaluation of longitudinal data

Author(s): Debra Wetcher-Hendricks, Moravian University

The McNemar test (1947) produces a chi-square value that, in contrast to the Pearson chi-square (1900) value, accounts for data from a single sample measured twice. Although extremely beneficial for data analysts, this test cannot be used for situations involving more than two measurements of the sample. For example, a longitudinal analysis that compares frequencies or percentages of a single sample divided into groups on three separate occasions or under three separate conditions cannot use McNemar's approach. The proposed repeated-measures chi-square design, however, compares frequencies or percentages from three or more dichotomous trials involving the same subjects. This process presented on this poster applies McNemar's basic principles to a multi-dimensional contingency table (Punnett Square) or to multiple contingency tables. A subsequent example demonstrates the formula's applicability in a three-trial situation.

Presenter: Shiying Xiao, University of Connecticut

Gaussian graphical models for functional connectivity analysis: A statistical review and applications to Alzheimer's disease data

Author(s): Shiying Xiao, University of Connecticut; W. Hudson Robb, Vanderbilt Memory and Alzheimer's Center; Jun Yan, University of Connecticut; Dandan Liu, Vanderbilt Memory and Alzheimer's Center/Vanderbilt University Medical Center; Panpan Zhang, Vanderbilt Memory and Alzheimer's Center/Vanderbilt University Medical Center

Functional connectivity analysis has emerged as a powerful tool for investigating the interactions among brain regions. In recent years, Gaussian graphical models (GGMs) have gained considerable attention for analyzing functional connectivity in brain imaging data. Nevertheless, a thorough investigation into their practical effectiveness has been lacking. In this paper, we present a comprehensive statistical review of GGMs and their application to Alzheimer's disease (AD) data. With the theoretical foundations of GGMs and their relevance to functional connectivity analysis as a backdrop, a fine-grained review of various estimation methods for GGMs is presented, including the graphical lasso (glasso), glasso with ridge penalty, graphical elastic net, adaptive glasso, SCAD, MCP, CLIME and TIGER. Following the review, we showcase the practical application of these methods to analyze AD data. Using the AD data from the Tennessee Alzheimer's Project (TAP), we illustrate and compare their utility in identifying alterations in brain networks associated with AD pathology. Additionally, we discuss challenges and future directions for applying GGMs to AD research, emphasizing their potential to uncover novel biomarkers for AD and other diseases.

Presenter: Wei Yang, Virginia Commonwealth University
Kernel-based partial Sufficient Dimension Reduction
Author(s): Wei Yang, Virginia Commonwealth University; Chenlu Ke, Virginia Commonwealth University

We propose a new partial Sufficient Dimension Reduction (SDR) method using a kernel-based regression sum of squares. Our method has the interpretability align with the linear regression model and is model-free. Compared with existing partial SDR methods, our method relaxes the requirements on model structure and predictors. Besides, our method is especially powerful for a categorical response, and can deal with a continuous response without slicing.

General Posters

Presenter: Ran Dai, University of Nebraska Medical Center
Controlling FDR in selecting group-level simultaneous signals from multiple data sources
Author(s): Runqiu Wang, University of Nebraska Medical Center; Ran Dai, University of Nebraska Medical Center; Hongying Dai, University of Nebraska Medical Center; Evan French, Virginia Commonwealth University; Cheng Zheng, University of Nebraska Medical Center

One challenge in exploratory association studies using observational data is that the associations between the predictors and the outcome are potentially weak and rare, and the candidate predictors have complex correlation structures. False discovery rate (FDR) controlling procedures can provide important statistical guarantees for replicability in predictor identification in exploratory research. In the recently established National COVID Collaborative Cohort (N3C), electronic health record (EHR) data on the same set of candidate predictors are independently collected in multiple different sites, offering opportunities to identify true associations by combining information from different sources. This paper presents a general knockoff-based variable selection algorithm to identify associations from unions of group-level conditional independence tests (simultaneous signals) with exact FDR control guarantees under finite sample settings. This algorithm can work with general regression settings, allowing heterogeneity of both the predictors and the outcomes across multiple data sources. We demonstrate the performance of this method with extensive numerical studies and an application to the N3C data.

Presenter: Run Fan, Vanderbilt University Medical Center
Statistical analysis of gene expression in subcutaneous adipose tissue and its association with ectopic lipid accumulation in HIV patients over time
Author(s): Run Fan, Vanderbilt University Medical Center; Oliver Zhao, Vanderbilt University; John Koethe, Vanderbilt University Medical Center; Fei Ye, Vanderbilt University Medical Center

Background: This study aims to identify genes expressed in subcutaneous adipose tissue (SAT) involved in adipocyte metabolism, inflammation, and immune function among Persons with HIV (PWH) on modern ART regimens.

Methods: Expression of 255 immune pathway genes and 77 genes related to adipocyte cellular regulation and lipid metabolism was measured in NanoString nCounter Plex panel. We employed linear mixed-effects models to assess the association between CT-derived measurements and gene expression over time (with a gene-by-time interaction term), adjusting for age, BMI, and diabetic group. Relationship between gene expression and tissue depots was shown

in network plot to visualize the degree of connectivity, the strength of significance, and the direction of interaction.

Results: Among 255 immune genes, preliminary analyses revealed significant gene-time interaction effects in 15 genes for skeletal muscle area and in 34 genes for skeletal muscle density. Among 77 adipocyte regulation genes, significant gene-time interactions were detected for skeletal muscle area in 4 genes and for skeletal muscle density in 9 genes. These data suggest that gene expression changes over time in PWH may contribute to the pathophysiology of skeletal muscle changes and underscore the potential for developing targeted interventions to reduce cardiometabolic disease risk in this population.

Presenter: Bo Ji, Boehringer Ingelheim Pharmaceuticals, Inc.
Statistical analysis in a Phase 2 dose finding trial for chronic kidney disease: A case study
Author(s): Bo Ji, Boehringer Ingelheim Pharmaceuticals, Inc.

In a phase 2 randomised, placebo-controlled dose finding study, the efficacy and safety of aldosterone synthase inhibition with and without empagliflozin, in patients with diabetic and non-diabetic chronic kidney disease was investigated. The trial employed a two-step randomisation process, stratified based on prognostic variables, to ensure balance between treatment groups.

The primary endpoint was the change in log-transformed UACR from baseline to week 14. The statistical analysis for this endpoint involved the use of a mixed model for repeated measures (MMRM), which accounted for within-subject variability and missing data. The Multiple Comparison Procedures – Modelling approach (MCP-Mod) was used to evaluate the dose–response relationship of the drug alone and combined with empagliflozin. This approach allowed for the simultaneous testing of multiple dose levels and the estimation of the optimal dose. The study also included secondary analyses of the proportion of patients achieving a 30% reduction in UACR, using logistic regression models adjusted for baseline UACR and treatment group.

The study provides valuable insights into the additive effects of aldosterone synthase inhibition when combined with empagliflozin. The results, including the statistically significant reduction in UACR with the optimal dose, could have significant implications for the treatment of chronic kidney disease.

Presenter: Hung-Chih Ku, DePaul University
Comparing genome-wide association studies from different statistical models
Author(s): Hung-Chih Ku, DePaul University; Zhengyang Zhou, University of North Texas Health Science Center; Chao Xing, University of Texas Southwestern Medical Center

Genome-wide association studies (GWAS) have served as primary methods for detecting associations between genetic variants and traits. Many statistical models have been applied for GWAS analysis, including linear, nonlinear, and logistic regression. In this study, we compare statistical models and evaluate validity and power of models in different settings. Based on these results, recommendations will be provided for the choice of models.

Presenter: Kuo-Jung Lee, National Cheng Kung University

Bayesian cumulative probit random effects model for longitudinal ordinal data using the hypersphere decomposition

Author(s): Kuo-Jung Lee, National Cheng Kung University; Ray-Bing Chen, National Cheng Kung University; Keunbaik Lee, Sungkyunkwan University

Longitudinal studies are prevalent in fields such as medicine, economics, and social sciences. This study focuses on analyzing longitudinal ordinal data, which often presents challenges due to serial correlations within subjects over time. We address this by introducing a Bayesian cumulative probit random effects model that accommodates both within-subject serial correlation and between-subject variance. Our model utilizes hypersphere decomposition to manage the constraints of positive definiteness and the high dimensionality in the correlation matrix. Additionally, we enhance computational efficiency through a hybrid Gibbs/Metropolis-Hastings algorithm, which expedites convergence in the Markov chain Monte Carlo (MCMC) simulations by generating cutoff points from truncated normal distributions. The effectiveness and robustness of our approach were validated through simulation studies under various scenarios, including complete data, missing completely at random (MCAR), and missing at random (MAR) conditions. We applied our model to two real-world datasets involving arthritis and lung cancer to demonstrate its practical utility.

Presenter: Handong Li, Cytel Inc.

Applying Bayesian rules for treatment selection in a seamless phase II/III design using cloud-based commercial software in combination with R code

Author(s): Handong Li, Cytel Inc.; Boaz Adler, Cytel Inc.; Valeria Mazzanti, Cytel Inc.; J. Kyle Wathen, Cytel Inc.

Background: Treatment selection is one of the pivotal aspects of seamless phase II/III designs. There are several accepted methods for selecting treatment groups to graduate to the phase III stage along with the control group.

Objective: Using cloud-based commercial simulation software in combination with R code, we applied custom Bayesian treatment selection rules to leverage prior data at the decision point for a novel phase II/III seamless design.

Methods: We inserted custom R code into design simulations to use Bayesian decision rules for the treatment selection at the interim analysis. We then conducted these simulations under different treatment effect assumptions using cloud-based simulation software to compare the operating characteristics of different design variations against likely execution scenarios.

Results: Our analysis shows that while commercial software offers a quick and reliable solution for standard designs, the integration of R code allows various approaches for analysis with greater flexibility.

Conclusion: The combined use of commercial software and personalized R code provides a powerful and flexible approach for applying different rules for treatment selection in seamless phase II/III designs, potentially leading to more efficient trial designs and improved outcomes.

Presenter: Austin Shih, Vanderbilt University

Power and sample size calculation for non-inferiority trials with treatment switching in intention-to-treat analysis with DRMST

Author(s): Austin Shih, Vanderbilt University; Chih-Yuan Hsu, Vanderbilt University Medical Center; Yu Shyr, Vanderbilt University Medical Center

Difference in Restricted Mean Survival Time (DRMST) has attracted attention and is increasingly used in non-inferiority trials. This is due to its often superior power in detecting treatment effects compared to Hazard Ratio (HR), even under a Proportional Hazards (PH) assumption. The non-parametric nature of DRMST also makes it ideal for instances where the PH assumption is not met. In non-inferiority trials, patient treatment switching under the PH assumption can lead to a violation of this assumption. This may result in underpowered trials when applying the widely used Intention-To-Treat (ITT) analysis. In this study, we propose a simulation-based approach, named nifts, to illustrate power reduction and calculate the necessary sample sizes for non-inferiority trials that allow treatment switching, based on DRMST. Our findings indicate that the switching probability and the switching time are key determinants in power and sample size calculation. The nifts tool is publicly available on GitHub for use.

Presenter: Chao Wang, US Food and Drug Administration

A multivariate equivalence test based on Mahalanobis distance with a data-driven margin

Author(s): Chao Wang, US Food and Drug Administration (FDA); Yu-Ting Weng, FDA; Shabo Liu, FDA; Tengfei Li, FDA; Meiyu Shen, FDA; Yi Tsong, FDA

Multivariate equivalence testing is needed in a variety of scenarios for drug development. For example, drug products obtained from natural sources may contain many components for which the individual effects and/or their interactions on clinical efficacy and safety cannot be completely characterized. Such lack of sufficient characterization poses a challenge for both generic drug developers to demonstrate and regulatory authorities to determine the sameness of a proposed generic product to its reference product. Another case is to ensure batch-to-batch consistency of naturally derived products containing a vast number of components, such as botanical products. The equivalence or sameness between products containing many components that cannot be individually evaluated needs to be studied in a holistic manner. Multivariate equivalence test based on Mahalanobis distance may be suitable to evaluate many variables holistically. Existing studies based on such method assumed either a predetermined constant margin, for which a consensus is difficult to achieve, or a margin derived from the data, where, however, the randomness is ignored during the testing. In this study, we propose a multivariate equivalence test based on Mahalanobis distance with a data-driven margin with the randomness in the margin considered. Several possible implementations are compared with existing approaches via extensive simulation studies.

Presenter: Yu-Ting Weng, US Food and Drug Administration

Impacts to IC50 estimation when concentration levels of hERG assay are reduced

Author(s): Yu-Ting Weng, US Food and Drug Administration; Dalong Huang, US Food and Drug Administration

The assessment of hERG safety margin is essential for estimating the risk for delayed repolarization and QT interval prolongation prior to first administration in humans. hERG safety

margin is usually defined by the half-inhibitory concentration (IC50) normalized to the drug's estimated clinical exposures. Per E14 and S7B Q&A, IC50 can be estimated using sufficient replicates and two or more concentrations achieving 20-80% block. In this project, we explored whether two concentrations achieving 20-80% block are sufficient to estimate an accurate IC50 and the potential impact of the two concentrations. We demonstrate that two concentrations achieving 20-80% block are insufficient to estimate an accurate IC50 and recommend alternative solutions.

Presenter: Zixuan Wu, University of Chicago

Causal mediation analysis for time-varying heritable risk factors with Mendelian Randomization

Author(s): Zixuan Wu, University of Chicago; Ethan Lewis, University of Chicago; Qingyuan Zhao, University of Cambridge; Jingshu Wang, University of Chicago

Understanding the causal pathogenic mechanisms of diseases is crucial in clinical research. When randomized controlled experiments are not available, Mendelian Randomization (MR) offers an alternative, leveraging genetic mutations as a natural "experiment" to mitigate environmental confoundings. However, most MR analyses treat the risk factors as static variables, potentially oversimplifying dynamic risk factor effects. The framework of life-course MR has been introduced to address this issue. However, current methods face challenges especially when the age-specific GWAS datasets have limited cohort sizes and there are substantial correlations between time points for a single trait. This study proposes a novel approach, estimating a unified system of structural equations for a sequence of temporally ordered heritable traits, requiring only GWAS summary statistics. The method facilitates statistical inference on direct, indirect, and path-wise causal effects and demonstrates superior efficiency and reliability, particularly with noisy GWAS data. By incorporating a spike-and-slab prior for genetic effects, the approach can address extreme polygenicity and weak instrument bias. Through this methodology, we uncovered a protective effect of BMI on breast cancer during a confined period of childhood development. We also analyzed how BMI, systolic blood pressure (SBP), and low-density cholesterol levels influence stroke risk across childhood and adulthood, and identified the intriguing relationships between these risk factors.

Presenter: Ao Yuan, Georgetown University

Doubly robust semiparametric model for causal inference with survival data

Author(s): Ao Yuan, Georgetown University; Tianmin Wu, Georgetown University; Ming Tan, Georgetown University

In observational studies in epidemiology and clinical trials often the treatment assignments are not completely random, and it is known that the naïve estimates of treatment effects may be biased, and causal inference methods are needed to get unbiased estimates. The doubly robust estimator (DRE) represents a significant advancement in causal inference. However, the literature on DRE with survival data is acutely scarce. The existing DRE methods for survival data are either complicated, based on optimal estimating equations, or based on Nelsen-Allen estimator. The former involves three components: the propensity score, the survival time, and the censoring time models. It is consistent if either the survival model or both the propensity score and censoring model are consistent, which is less than being doubly robust. The latter is doubly robust, but is based on the less popular Nelsen-Allen estimator. To address these issues,

we propose a new robust estimate utilizing the most commonly used Kaplan-Meier estimator and Stute representation with a weighted empirical form. The proposed estimator involves only two components: the propensity score model and the conditional survival model. It is doubly robust as it only requires one of the two models to be correctly specified as opposed to requiring two of the three models to be correctly specified. The new robust estimator thus has a much simpler structure than the existing ones. In addition, we use a semiparametric model for the propensity score, which is shown to further increase the robustness. The asymptotic behavior is derived, and simulation studies are conducted to evaluate the finite sample performance of the proposed semiparametric DRE.

Presenter: Zihan Zhu, University of Arizona

Bayesian Poisson regression with spatially dependent global-local shrinkage prior

Author(s): Zihan Zhu, Xueying Tang, Shuang Zhou

For hurricane prediction problems, the challenge often lies in the small sample sizes and high-dimensional, spatially correlated covariates. Traditional regularized regression models, such as Poisson regression with an elastic net penalty, struggle with these data characteristics, leading to bad predictions. We propose a Bayesian Poisson regression model that incorporates spatially dependent global-local shrinkage priors. It is designed to discern subtle signals within global field data and simultaneously generate predictions. Our model employs a Conditional Autoregressive (CAR) Gaussian prior for the spatially dependent covariates to account for spatial correlations. Additionally, it applies global-local shrinkage factors to the CAR prior to effectively mitigate the influence of inactive regions while maintaining the significance of active ones, which also helps to decouple the correlation effects between active and inactive regions. The shrinkage factors themselves are assigned with half-Cauchy and log-Cauchy priors, with the latter demonstrating superior performance in scenarios of weak signals and strong spatial correlation among covariates. A Metropolis-within-Gibbs sampler is developed for computational implementation. Simulation studies affirm the efficacy of the proposed model. When applied to the North Atlantic hurricane prediction problem, our model outperforms both the elastic net approach and traditional climatology, closely rivaling the University of Arizona's (UA) model, which serves as the benchmark "oracle" in this context.

Tuesday, June 18, 2024

7:30 am – 6:30 pm	Registration	Symphony Foyer Entrance (Lobby Level)
7:30 – 8:30 am	Breakfast	Symphony Foyer (Lobby Level)
8:30 – 9:30 am	Keynote: Yu Shen, PhD, FASA	Symphony 2&3 (Lobby Level)
9:30 – 10:00 am	Coffee Break	Symphony Foyer (Lobby Level)
10:00 – 11:30 am	Invited Sessions & Panels	
11:30 am – 1:00 pm	Lunch	
1:00 – 2:30 pm	Invited Sessions	
2:30 – 3:00 pm	Coffee Break	Symphony Foyer (Lobby Level)
3:00 – 4:30 pm	Invited Sessions	
4:30 – 5:00 pm	Coffee Break	Symphony Foyer (Lobby Level)
5:00 – 6:30 pm	Invited Sessions	
7:00 – 10:00 pm	Banquet and Student Awards Ceremony, with keynote by Mingyao Li, PhD, FASA, FAAAS (tickets required)	Symphony 2&3 (Lobby Level)

Online agenda: symposium2024.icsa.org/detailed-agenda

Ideas for lunch: symposium2024.icsa.org/nashville

Share your experience on social media! #ICSAnashville

Keynote Talk – Tuesday, June 18

Synergizing multiple data sources: Enhancing precision medicine risk prediction

Yu Shen, PhD, FASA

8:30 am, Symphony 2&3 (Lobby Level)

Advances in data science and machine learning present both challenges and significant opportunities for statisticians to contribute to precision medicine by harnessing multiple data sources. Potential benefits include improved statistical efficiency and enhanced accuracy in risk prediction, particularly for rare types of cancer, as well as a better understanding of the natural history of chronic diseases. However, such integration of data also poses statistical challenges due to potential incomparability between different data sources and variations in sampling mechanisms.

In this talk, we discuss recent developments in the field and describe adaptive estimation procedures that utilize combined information to assess the degree of information borrowing from external data, whether aggregated or at individual level. We also explore various statistical models for integrating data from multiple cohorts under different sampling schemes. These advancements not only enhance predictive models but also address issues related to data heterogeneity, ultimately contributing to more reliable evidence in precision medicine.

Dr. Yu Shen is Professor and Interim Chair of Biostatistics at the University of Texas MD Anderson (MDA) Cancer Center, where she holds the Conversation with a Living Legend Professorship. She obtained her PhD in biostatistics from the University of Washington and has been a faculty member of MDA since then. Her research spans the areas of novel biostatistics methodology research, health service research in cancer early detection, and personalized cancer treatment. Her priority is to develop state-of-the-art statistical methods and models to address important clinical research questions. She has developed statistical methods in the areas of data integration and adaptive clinical trial designs, as well as modeling the natural history of cancer and survival data subject to biased sampling, with continuous NIH grant support for her statistical method development and health economic research. Dr. Shen has held leadership roles in the ASA and other statistical societies, including Program Chair for the Lifetime Data Analysis (LIDA) Interest Group and Secretary/Treasurer of the Biometrics Section, as well as serving on the ENAR Program Committee and ICSA Board of Directors. She has also served as an associate editor for four major biostatistics journals. She has been recognized for her contributions to the field with a Cancer Research and Prevention Foundation award, as well as being named a finalist for the Julie and Ben Rogers Award for Excellence and elected ASA Fellow.

P3: APPLICATIONS AND ADVANCES WITH LARGE LANGUAGE MODELS IN BIOPHARMACEUTICAL STATISTICS

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Symphony 1 (Lobby Level)

Organized by Arinjita Bhattacharyya, Associate Principal Scientist at Merck, and moderated by Hongtu Zhu, Professor of Biostatistics at University of North Carolina, this panel of LLM experts will share their observations and insights on what's new and exciting in the realm of deep learning. The panelists are:

- Margaret Gamalo, Vice President and Statistics Therapeutic Area Head, Pfizer
- Li Wang, Head of Statistical Innovation, AbbVie
- Thomas Jemielita, Principal Scientist, Merck
- Hulin Wu, Professor of Bioinformatics and Systems Medicine, University of Texas Health Science Center
- Meng Hu, Lead Chemical Engineer, FDA

S31: MODERN STATISTICAL LEARNING ADVANCES IN GENETICS

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Blackbird A (Mezzanine Level)

Organizer: Tianying Wang, Colorado State University
Chair: Tianying Wang, Colorado State University

10:00 AM–10:20 AM Speaker: Hongyu Zhao, Yale University

Genetic risk predictions across populations

Author(s): Hongyu Zhao, Yale University

The disparity in genetic risk prediction accuracy between European and non-European individuals highlights a critical challenge in health inequality. To bridge this gap, we introduce JointPRS, a novel method that models multiple populations jointly to improve genetic risk predictions for non-European individuals. JointPRS has three key features. First, it encompasses all diverse populations to improve prediction accuracy, rather than relying solely on the target population with a singular auxiliary European group. Second, it autonomously estimates and leverages chromosome-wise cross-population genetic correlations to infer the effect sizes of genetic variants. Lastly, it provides an auto version that has comparable performance to the tuning version to accommodate the situation with no validation dataset. Through extensive simulations and real data applications to 22 quantitative traits and four binary traits in East Asian, nine quantitative traits and one binary trait in African, and four quantitative traits in South Asian populations, we demonstrate that JointPRS outperforms state-of-art methods, improving the prediction accuracy for both quantitative and binary traits in non-European populations. This is joint work with Leqi Xu, Geyu Zhou, Wei Jiang, and Leying Guan.

10:20 AM–10:20 AM Speaker: Haoyu Zhang, National Cancer Institute, NIH

A novel approach for identifying genetic associations with heterogeneous cancer subtypes risk

Author(s): Sheng Fu, National Cancer Institute; Kai Yu, National Cancer Institute; Haoyu Zhang, National Cancer Institute

Breast cancer is a complex disease with diverse molecular subtypes, each with distinct etiologies, clinical presentations, and outcomes. Detecting the association between disease subtypes and common germline variants is a challenging task due to the disease's heterogeneity. Existing methods are less effective in the presence of interaction effects, and they are computationally intensive. To address this issue, we propose a novel model named TOPO, that efficiently detects variants exhibiting subtype heterogeneity. Our comprehensive test procedure combines three different model structures, including fixed-effect and random-effect two-stage polytomous models, and uses a fast and powerful procedure to combine three p-values. Through extensive simulation studies, we show that TOPO has well controlled type-one-error, and superior performance in statistical power and computational time. We apply TOPO to the up-to-date largest genome-wide association study of 138,209 breast cancer cases and 121,663 controls from the Breast Cancer Association Consortium. After filtering out known risk loci, we identified eight novel variants (p-value < 5 x 10-8). Our findings highlight the importance of considering tumor heterogeneity in identifying new loci, enhancing our understanding of breast cancer's etiologic heterogeneity, and informing subtype-specific genetic scores for precision prevention.

10:40 AM–11:00 AM Speaker: Hae Kyung Im, University of Chicago

Leveraging deep learning models to enable prediction of omics data in low power settings
Author(s): Yichao Zhou, University of Chicago; Mengjie Chen, University of Chicago; Ravi Madduri, Argonne National Laboratory; Temidayo Adeluwa, University of Chicago; Hae Kyung Im, University of Chicago

Genetic prediction of molecular traits are powerful ways to make sense of GWAS discoveries as demonstrated by TWAS and related methods. However, detecting and predicting disease relevant regulatory effects can be difficult due to their cell type specificity and the cost to assay large number of samples. I will present a novel method, scPrediXcan, that leverages deep learning methods to predict cell type specific expression directly from DNA sequences. Application to type 2 diabetes shows that our method outperforms traditional population based methods and enables prediction model training in low power settings.

11:00 AM–11:20 AM Speaker: Lin Hou, Tsinghua University

Inference of heterogeneous effect in Perturb-Seq experiments
Author(s): Zichu Fu, Lin Hou

The integration of CRISPR screening and single-cell RNA sequencing has arisen as a powerful tool for profiling the impact of genetic perturbations on the entire transcriptome at the single-cell scale. Computational methods have been developed to estimate average perturbation effects. Here we present a new method that disentangles perturbation effect from heterogeneous cell state, and infers perturbation effect at single-cell resolution. We demonstrate in simulation studies and real datasets that our method facilitates genetic interaction analysis, clustering of perturbation effect, and prioritization of genes in various biological processes.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S32: NEW MODELS AND INFERENTIAL METHODS FOR COMPLICATED BIOMEDICAL DATA

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Lyric (Lobby Level)

Organizer: Min Zhang, Tsinghua University
Chair: Kevin (Zhi) He, University of Michigan

10:00 AM–10:20 AM Speaker: Zhengwu Zhang, University of North Carolina at Chapel Hill

Alignment of continuous brain connectivity

Author(s): Zhengwu Zhang, University of North Carolina at Chapel Hill

We address the brain network registration problem in this talk. The brain network is commonly represented by an adjacency matrix, A, with VxV elements, where V indicates the number of nodes. The nodes are brain regions, which are assumed to be aligned across individuals. However, existing alignment methods rarely consider the connectivity information, partially because under the adjacency matrix representation, brain network alignment becomes a computationally demanding node-matching problem. Without properly aligned nodes, edge variability will be inflated, causing reduced statistical power in downstream network analysis tasks. Here, we introduce an efficient brain network registration based on a novel network representation, continuous connectivity (ConCon) that describes connectivity patterns between any pair of points on the brain's cortical surface. Under the ConCon representation, network registration becomes a task of finding optimal diffeomorphisms to match cortical surfaces based on their ConCon profiles. We develop an efficient optimization algorithm for this purpose. Our analyses using both the Human Connectome Project (HCP) and Adolescent Brain Cognitive Development (ABCD) study datasets show that our method significantly improves current connectivity connectome tasks.

10:20 AM–10:20 AM Speaker: Zhimei Ren, University of Pennsylvania

Conformalized survival analysis

Author(s): Emmanuel Candès, Stanford University; Lihua Lei, Stanford University; Zhimei Ren, University of Pennsylvania; Yu Gui, University of Chicago; Rohan Hore, University of Chicago; Rina Barber, University of Chicago

Existing survival analysis techniques heavily rely on strong modeling assumptions and are, therefore, prone to model misspecification errors. In this talk, I will introduce inferential methods based on ideas from conformal prediction, which can wrap around any survival prediction algorithm to produce calibrated, covariate-dependent lower predictive bounds on survival times. In the Type I right-censoring setting, when the censoring times are completely exogenous, the lower predictive bounds have guaranteed coverage in finite samples without any assumptions other than that of operating on independent and identically distributed data points. Under a more general conditionally independent censoring assumption, the bounds satisfy a doubly robust property which states the following: marginal coverage is approximately guaranteed if either the censoring mechanism or the conditional survival function is estimated well. Further, we demonstrate that the lower predictive bounds remain valid and informative for other types of censoring. The validity and efficiency of our procedure are demonstrated on synthetic data

and real data.

10:40 AM–11:00 AM Speaker: Panpan Zhang, Vanderbilt University Medical Center

An anchoring-event-based sigmoidal mixed model with application to Alzheimer's disease progression

Author(s): Kaidi Kang, Vanderbilt University Medical Center; Panpan Zhang, Vanderbilt University Medical Center; Timothy Hohman, Vanderbilt University Medical Center; Dandan Liu, Vanderbilt University Medical Center

Sigmoidal mixed models are prevalent for characterizing longitudinal trajectories of biomarkers related to chronic neurodegenerative disorders like Alzheimer's disease (AD). This presentation will address two major challenges arising from longitudinal data analysis. The first challenge is about identifying an appropriate and broadly applicable time scale, and the second challenge is about selection bias caused by filtering part of cohort participants due to model limitations. In response to these two challenges, we proposed an anchoring-event-based sigmoidal mixed model. The proposed model is evidenced to outperform competing methods through simulations. The model is also applied to a multi-center AD progression study.

11:00 AM–11:20 AM Speaker: Shizhe Chen, University of California, Davis

Simultaneous clustering and estimation of additive shape invariant models for recurrent event data

Author(s): Zitong Zhang, University of California, Davis; Shizhe Chen, University of California, Davis

Technological advancements have enabled the recording of spiking activities from large neuron ensembles, presenting an exciting yet challenging opportunity for statistical analysis. This project considers the challenges from a common type of neuroscience experiments, where randomized interventions are applied over the course of each trial. The objective is to identify groups of neurons with unique stimulation responses and estimate these responses. The observed data, however, comprise superpositions of neural responses to all stimuli, which is further complicated by varying firing latencies across neurons. We introduce a novel additive shape invariant model that is capable of simultaneously accommodating multiple clusters, additive components, and unknown time-shifts. We establish conditions for the identifiability of model parameters, offering guidance for the design of future experiments. We examine the properties of the proposed algorithm through simulation studies and apply the proposed method on neural data collected in mice.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S33: ANALYTICAL ADVANCES IN EMERGING HIGH-DIMENSIONAL BIOMEDICAL APPLICATIONS

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Ocean Way (Mezzanine Level)

Organizer: Yize Zhao, Yale University
Chair: Simiao Gao, Yale University

10:00 AM–10:20 AM Speaker: Lin Zhang, University of Minnesota

Statistical image partitioning methods for lesion-wise cancer detection of prostate MRI

Author(s): Maria Masotti, University of Michigan; Lin Zhang, University of Minnesota; Ethan Leng, ; Greg Metzger, University of Minnesota; Joe Koopmeiners, University of Minnesota

Imaging plays an important role in cancer diagnosis and staging by noninvasively evaluating the presence and extent of local and distant disease. Computer aided detection algorithms are being developed for fast and reproducible cancer diagnosis from complex and high-dimensional medical imaging data. While extensive statistical methods have been developed for voxel-wise cancer classification, existing lesion segmentation methods primarily rely on deep learning methods centered at the convolutional neural networks. We have developed novel statistical image partitioning methods for lesion-wise cancer detection using imaging data, which jointly estimate the lesion boundaries and the spatial processes within each partitioned region in a Bayesian framework. We show through simulations and application to prostate cancer imaging data that the proposed methods well estimate the number and boundaries of cancerous regions of arbitrary number and shape with higher sensitivity and specificity compared to competitive methods.

10:20 AM–10:20 AM Speaker: Hengrui Luo, Rice University

Reducing the dimension of single-cell data: Multi-group structures and beyond

Author(s): Hengrui Luo, Rice University; Didong Li, University of North Carolina, Chapel Hill

In single-cell data analysis, recent progress has centered on dimension reduction, leveraging group structures and periodicities. Advanced statistical methods are utilized to identify unique structures in distinct groups, aiding in distinguishing between healthy and diseased cells. This is key for uncovering complex cellular behaviors and understanding disease mechanisms. Concurrently, there's a shift toward systematic hyperparameter tuning in dimension reduction algorithms, enhancing the reproducibility of visualizations in complex biological data. These advancements mark significant strides in statistical methodologies for single-cell data analysis, emphasizing the need for precision and sophisticated algorithms. The combined focus on group-specific analysis and optimization techniques underscores the complexities and challenges of performing dimension reduction in high-dimensional biological data efficiently.

10:40 AM–11:00 AM Speaker: Yi Zhao, Indiana University

Beyond massive univariate tests: Covariance regression reveals complex patterns of brain functional connectivity

Author(s): Yi Zhao, Indiana University

Studies of brain functional connectivity typically involve massive univariate tests, performing statistical analysis on each individual connection. In this study, we consider the problem of regressing covariance matrices on associated covariates. The goal is to use covariates to explain variation in covariance matrices across units. As such, we introduce Covariate Assisted Principal (CAP) regression, an optimization-based method for identifying components associated with the covariates using a generalized linear model approach. For high-dimensional data, a well-conditioned linear shrinkage estimator of the covariance matrix is introduced. With multiple covariance matrices, the shrinkage coefficients are proposed to be common across matrices. Theoretical studies demonstrate that the proposed covariance matrix estimator is optimal achieving the uniformly minimum quadratic loss asymptotically among all linear combinations of the identity matrix and the sample covariance matrix. Under regularity conditions, the proposed estimator of the model parameters is consistent. We develop computationally efficient algorithms to jointly search for common linear projections of the covariance matrices, as well as the regression coefficients. The superior performance of the proposed approach over existing methods is illustrated through simulation studies. Applied to resting-state functional magnetic resonance imaging (fMRI) studies, the proposed approach regresses whole-brain functional connectivity on covariates and enables the identification of relevant brain subnetworks.

11:00 AM–11:20 AM Discussant: Yimei Li, St. Jude Children's Research Hospital

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S34: RECENT ADVANCEMENTS IN STATISTICAL METHODS FOR COMPLEX LIFETIME DATA

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Platinum (Lower Level)

Organizer: Jing Ning, University of Texas MD Anderson Cancer Center
Chair: Jin Piao, University of Southern California

10:00 AM–10:20 AM Speaker: Ming Wang, Case Western Reserve University

Joint hierarchical modeling of recurrent and terminal events with dynamic predictions

Author(s): Ming Wang, Case Western Reserve University; Yifan Yang, Case Western Reserve University; Chixiang Chen, University of Maryland

In clinical and observational studies, recurrent events are frequently encountered, including repeated episodes of hospitalizations, strokes, and others. Throughout the follow-up period, these recurrent events may be censored by a competing risk event (e.g., death), resulting in dependent censoring. While joint modeling of recurrent events with the terminal event has garnered considerable attention, further advancements in analytical methods are necessary to meet real-world application needs. One commonly utilized approach is joint frailty modeling with a shared frailty for recurrent and terminal event processes. However, there are limitations, such as the strong assumption of conditional independence, which can be easily violated in practice. Also, handling multiple types of recurrent events, such as repeated occurrences of heart failure, myocardial infarction, stroke, or other cardiovascular diseases, is crucial but has not been thoroughly explored. Here, we will provide an overview of existing methods and discuss recently proposed ones to address these gaps, highlighting their advantages and disadvantages. Furthermore, evaluating patient prognosis and dynamically predicting the risk of death are clinically significant, and utilizing developed models considering historical recurrent events is expected to enhance medical decisions and improve healthcare outcomes. These methods will be illustrated using large-scale real data. Finally, we will discuss current methodological gaps and propose directions for future research.

10:20 AM–10:20 AM Speaker: Sy Han Chiou, Southern Methodist University

Regression analysis of bivariate survival data using pseudo-observations

Author(s): Sy Han Chiou, Southern Methodist University; Chien-Lin Su, Worldwide Clinical Trials; Feng-Chang Lin, University of North Carolina

Copula models have become increasingly popular in various fields as they provide effective tools for modeling correlated responses. In modeling multivariate survival data, copula models offer flexibility by enabling users to specify both the marginal survival functions and the association structure between them. In this study, we consider a semiparametric transformation model to define the marginal survival functions and a conditional Archimedean copula to address the associations among different types of survival times. To expedite computation, we introduce pseudo-observations for both the marginal survival and association components and implement inference using generalized estimating equation techniques. Additionally, we explore variable selection and goodness-of-fit tests to aid in the selection of appropriate copula models. The effectiveness of our proposed methods is demonstrated through extensive simulations.

10:40 AM–11:00 AM Speaker: Jian Wang, University of Texas MD Anderson Cancer Center

Assessing dynamic and predictive discrimination of recurrent event models using a time-dependent C-index

Author(s): Jian Wang, University of Texas MD Anderson Cancer Center; Xinyang Jiang, University of Texas MD Anderson Cancer Center; Jing Ning, University of Texas MD Anderson Cancer Center

In recent decades, there has been a growing interest in the analysis of recurrent event data. A crucial aspect of developing a risk prediction model for such data is accurately identifying individuals with varying risks of experiencing a recurrent event. While the concordance index (C-index) effectively assesses the overall discriminative ability of a regression model for recurrent event data, there is a need for a local measure to capture the model's dynamic performance over time. In this study, we propose a time-dependent C-index measure to evaluate the model's discriminative ability locally. We formulated this measure as a function of time using a flexible parametric model and constructed a concordance-based likelihood for estimation and inference. A perturbation-resampling procedure was employed to estimate variance. We conducted extensive simulations to investigate the finite-sample performance of the proposed time-dependent C-index and its estimation procedure. The proposed time-dependent C-index was applied to a study of re-hospitalization among patients with colorectal cancer to assess the discriminative capability of different models.

11:00 AM–11:20 AM Speaker: Shikun Wang, Columbia University

A general backward joint model of longitudinal and survival data with application to dynamic prediction

Author(s): Shikun Wang, Columbia University; Liang Li, University of Texas MD Anderson Cancer Center

In the dynamic prediction (DP) of time-to-event outcomes of subjects using the longitudinal medical history, a key step is to jointly model the distribution of time-to-event and longitudinal data. The widely used forward joint model (FJM) with shared random effects has been widely applied to the dynamic prediction problems, however, it can be computationally intensive as the number of longitudinal predictors increases. The backward joint model (BJM) is a numerically tractable and stable solution to dynamic prediction problems with multiple longitudinal responses. Here, we propose a general BJM and its analytic relationships with FJM under conditions that allows more accessible estimators for sophisticated techniques. We illustrate the model using simulations and a data example.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S35: METHODS FOR BIOMARKERS IN BIOMEDICAL RESEARCH

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Sound Emporium A/B (Mezzanine Level)

Organizer: Aiyi Liu, National Institute of Child Health and Human Development (NICHD), NIH
Chair: Zhen Chen, National Institute of Child Health and Human Development (NICHD), NIH

10:00 AM–10:20 AM Speaker: Guoqing Diao, The George Washington University

Improving power in adaptive expansion of biomarker populations in phase 3 clinical trials

Author(s): Guoqing Diao, The George Washington University; Xun Jiang, Amgen; Donglin Zeng, University of Michigan; May Mo, Amgen; Amy Xia, Amgen; Joseph Ibrahim, University of North Carolina at Chapel Hill

With the availability of unprecedented human genomic biomarker data, incorporating such biomarker data has received a lot of attention in phase 3 clinical trials. One particular enrichment design proposed recently in the literature is to recruit more biomarker positive patients in an all-comer study if the treatment effect in the biomarker negative group is less promising than expected. The intuition is to improve the chance of success of the trial since the success rate in the all-comer population may be low. We propose an enrichment design that unifies the existing biomarker adaptive designs for phase 3 clinical trials. In addition, we propose a new test accounting for the correlations among the test statistics based on different groups of patients, including all-comers, biomarker positive patients only, and biomarker negative patients only. We investigate the theoretical properties of the design and demonstrate the new test accurately controls the type I error rate and gains power over existing methods through extensive simulations. A computer program is developed for power calculations given a set of design parameters, including the proportion of biomarker positive patients, the distribution of the failure time in each treatment and biomarker group, and the number of patients in the first stage and the second stage (i.e., the enrichment stage), among others.

10:20 AM–10:20 AM Speaker: Zhiwei Zhang, Gilead Sciences, Inc.

Efficient estimation strategies for biomarker studies embedded in randomized trials

Author(s): Wei Zhang, Chinese Academy of Sciences; Zhiwei Zhang, Gilead Sciences, Inc.; James Troendle, National Institutes of Health; Aiyi Liu, National Institutes of Health

Predictive and prognostic biomarkers are increasingly important in clinical research and practice. Biomarker studies are frequently embedded in randomized clinical trials with biospecimens collected at baseline and assayed for biomarkers, either in real time or retrospectively. Here we propose efficient estimation strategies for two study settings in terms of biomarker ascertainment: the complete-data setting in which the biomarker is measured for all subjects in the trial, and a two-phase sampling design in which the biomarker is measured retrospectively for a random subsample of subjects selected in an outcome-dependent fashion. In both settings, efficient estimating functions are characterized using semiparametric theory and approximated using data-adaptive machine learning methods, leading to estimators that are consistent, asymptotically normal, and (approximately) efficient under general conditions. The proposed methods are evaluated in simulation studies and applied to real data from a cancer trial.

10:40 AM–11:00 AM Speaker: Pang Du, Virginia Tech

Likelihood ratio combination of multiple biomarkers via smoothing spline estimated densities
Author(s): Zhiyuan Du, Virginia Tech; Pang Du, Virginia Tech; Aiyi Liu, NICHD

The diagnostic accuracy of multiple biomarkers in medical research is crucial for detecting diseases and predicting patient outcomes. An optimal method for combining these biomarkers is essential to maximize the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). Although the optimality of the likelihood ratio has been proven by Neyman and Pearson, challenges persist in estimating the likelihood ratio, primarily due to the estimation of multivariate density functions. In this study, we propose a non-parametric approach for estimating multivariate density functions by utilizing Smoothing Spline density estimation to approximate the full likelihood function for both diseased and non-diseased groups, which compose the likelihood ratio. Simulation results demonstrate the efficiency of our method compared to other biomarker combination techniques under various settings for generated biomarker values. Additionally, we apply the proposed method to a real-world study aimed at detecting childhood autism spectrum disorder (ASD), showcasing its practical relevance and potential for future applications in medical research.

11:00 AM–11:20 AM Speaker: Aiyi Liu, NICHD

OUCopula: Bi-channel multi-label copula-enhanced adapter-based CNN for myopia screening based on OU-UWF images
Author(s): Yang Li, Fudan University; Qiuyi Huang, Hong Kong Polytechnic University; Chong Zhong, Hong Kong Polytechnic University; Danjuan Yang, Fudan University; Meiyan Li, Fudan University; A.H. Welsh, Australian National University; Aiyi Liu, NICHD; Bo Fu, Fudan University; Catherine C. Liu, Hong Kong Polytechnic University; Xingtao Zhou, Fudan University

Myopia screening using cutting-edge ultra-widefield (UWF) fundus imaging is potentially significant for ophthalmic outcomes. Current multidisciplinary research between ophthalmology and deep learning (DL) concentrates primarily on disease classification and diagnosis using single-eye images, largely ignoring joint modeling and prediction for Oculus Uterque (OU, both eyes). Inspired by the complex relationships between OU and the high correlation between the (continuous) outcome labels (Spherical Equivalent and Axial Length), we propose a framework of copula-enhanced adapter convolutional neural network (CNN) learning with OU UWF fundus images (OUCopula) for joint prediction of multiple clinical scores. We design a novel bi-channel multi-label CNN which can (1) take bichannel image inputs subject to both high correlation and heterogeneity (by sharing the same backbone network and employing adapters to parameterize the channel-wise discrepancy), and (2) incorporate correlation information between continuous output labels (using a copula). Solid experiments show that OUCopula achieves satisfactory performance in myopia score prediction compared to backbone models. Moreover, OUCopula can far exceed the performance of models constructed for single-eye inputs. Importantly, our study also hints at the potential extension of the bi-channel model to a multi-channel paradigm and the generalizability of OUCopula across various backbone CNNs.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S36: NEW METHODS IN BIOMEDICAL DATA ANALYSIS WITH THE APPLICATIONS TO GENOMICS STUDIES

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Southern Ground A/B (Mezzanine Level)

Organizer: Yichuan Zhao, Georgia State University
Chair: Zhigang Li, University of Florida

10:00 AM–10:20 AM Speaker: Susmita Datta, University of Florida

Understanding cell microenvironments from spatially resolved transcriptomics data

Author(s): Dongyuan Wu, University of Florida; Jeremy Gaskins, University of Louisville; Michael Sekula, University of Louisville; Susmita Datta, University of Florida

Understanding cellular microenvironment such as cellular communication through biochemical signaling is fundamental to every biological activity. Investigating cell signaling diffusions across cell types can further help understand biological mechanisms. In recent years, this has become an important research topic using single-cell sequencing and specifically spatially resolved transcriptomics data. As far as we know, most existing methods focus on providing an ad hoc measurement to estimate intercellular communication instead of relying on a statistical model. It is undeniable that those descriptive statistics are straightforward and accessible, but a suitable statistical model can provide more accurate and reliable inference. In this research, we propose a generalized linear regression model to infer cellular communications from spatially resolved transcriptomics data, especially spot-based data. Our BAyesian Tweedie modeling of COMmunications (BATCOM) method estimates the communication scores between cell types with the consideration of their corresponding distances.

10:20 AM–10:20 AM Speaker: Dongmei Li, University of Rochester Medical Center

scMAE: A deep-learning algorithm for single-cell RNA sequencing differential analysis

Author(s): Dongmei Li, University of Rochester Medical Center; Pinxin Liu, University of Rochester; Shijian Deng, University of Texas at Dallas; Zidian Xie, University of Rochester Medical Center

Differential gene expression (DGE) analysis stands as a crucial step in the single-cell RNA sequencing (scRNA-seq) data analysis pipeline, offering insights into novel cell types and gene signatures contributing to cellular heterogeneity. While numerous statistical methods exist for DGE analysis in scRNA-seq data, none explicitly consider gene-gene interactions within this context. In response, we introduce scMAE, a novel approach leveraging Mask AutoEncoder (MAE), to enhance DGE analysis in scRNA-seq data by incorporating gene-gene interactions. scMAE is a deep-learning algorithm rooted in transformer architecture, commonly employed in Natural Language Processing (NLP) and Computer Vision (CV). Through simulation studies, scMAE demonstrates superior performance compared to many existing statistical methods for DGE analysis in scRNA-seq data.

10:40 AM–11:00 AM Speaker: Qi Long, University of Pennsylvania

Accounting for network noise in graph-guided Bayesian modeling of high-dimensional omics data

Author(s): Wenrui Li, University of Pennsylvania; Changgee Chang, Indiana University; Suprateek Kundu,

University of Texas MD Anderson Cancer Center; Qi Long, University of Pennsylvania

High-dimensional omics data offer great promise in advancing precision medicine. Knowledge-guided statistical methods for analysis of omics data that can incorporate biological knowledge represented by graphs such as functional genomics have been shown to improve variable selection and predication accuracy and yield biologically more interpretable results, they typically use biological graph extracted from existing databases which is known to be incomplete and contain false edges. To address this issue, we propose a new knowledge-guided Bayesian modeling framework that treats the true biological graph as unknown or latent. Our model uses an adaptive structured shrinkage prior to incorporate the latent true biological graph to facilitate variable selection, and another set of priors motivated by the latent scale network model to connect two sources of noise-contaminated graph data—namely, biological graph extracted from a database and estimated covariance matrix for covariates, to the latent true graph. We develop an efficient MCMC algorithm for posterior sampling. We demonstrate the advantages of our model in simulations, and analysis of an AD genomics dataset.

11:00 AM–11:20 AM Speaker: Meredith Ray, University of Memphis

Using a novel non-parametric clustering approach to identifying patterns of multi-genetic/epigenetic factors

Author(s): Meredith Ray, University of Memphis; Kip Handwerker, University of Memphis; Lauren Sobral, Indigo; Allison Plaxco, University of Memphis; Hongmei Zhang, University of Memphis

As technology continues to rapidly advance, the ability to generate high-dimensional and high throughput data also advances. With accompanying improved computation power, the need for improved data mining methods is dire. We propose a novel clustering approach for multi-dimensional and mixed type of data. This was built with genetic and epigenetic data including single nucleotide polymorphisms (SNPs), DNA methylation, and gene expression for a set of genes as the motivation. The approach follows the typical K-means framework but have formulated a novel Euclidean-distance-based metric to assess distances between clustering objects; thus, considering complex joint effects of SNPs and DNA methylation on the expression of a gene. Simulations studies were conducted and demonstrated high sensitivity, specificity, and accuracy with respect to cluster assignment. As genetic/epigenetic data was the motivation, we applied the method to a data set from a birth cohort on the Isle of Wight, UK (data includes SNPs, DNA methylation, and gene expressions) and aimed to identify clusters and tested if those clusters were associated with eczema, possibly identifying eczema-related genes and examine genetic and epigenetic patterns.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S37: MODERN AND INNOVATIVE METHODS IN ANALYZING COMPLEX MEDICAL DATA

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Blackbird B (Mezzanine Level)

Organizer: Esra Kurum, University of California, Riverside
Chair: Esra Kurum, University of California, Riverside

10:00 AM–10:20 AM Speaker: Saman Muthukumarana, University of Manitoba

Discovering long COVID symptom patterns and prediction using clinical notes data

Author(s): Saman Muthukumarana, University of Manitoba

In this talk, I will discuss methods for understanding the patterns and behaviour of long COVID symptoms reported by patients on the Twitter social media platform, which is vital in identifying early frequent symptoms and establish relationships between symptoms among long COVID patients. I will then discuss machine learning classification models adept at unravelling the intricacies of distinguishing long COVID cases from many other health conditions. Using natural language processing techniques, we outline a confirmed long COVID group to serve as the foundation for classification. The aim is to establish a precise and dependable method for identifying long COVID patients, achieved through a comprehensive assessment of classification and re-sampling techniques. This includes the integration of patient attributes and both pre-and post-COVID symptoms linked to Long Covid Syndrome (LCS) as input variables.

10:20 AM–10:20 AM Speaker: Zhaoxia Yu, University of California, Irvine

A robust latent-variable method for clustered data, with applications to multi-subject single-cell omics data

Author(s): Mingyu Du, UC Irvine; Kevin Johnston, UC Irvine; Veronica Berrocal, UC Irvine; Wei Li, UC Irvine; Eran Mukamel, University of California San Diego; Xiangmin Xu, UC Irvine; Zhaoxia Yu, UC Irvine

Technological advancements have greatly enhanced our understanding of biological processes by allowing us to quantify key biological metrics such as gene expression, protein levels, and microbiome compositions. In particular, these breakthroughs now enabled us to achieve increasingly higher levels of resolution in our measurements, exemplified by our ability to comprehensively profile biological information at the single-cell level. However, the analysis of such data faces several critical challenges: limited sample sizes, non-normal distributions with potential dropouts and outliers, and repeated measurements from the same biological units. Existing methods often focus on addressing one or two of the issues. In this article, we propose a novel U-statistic based Latent Variable (ULV) model that takes advantage of the robustness of rank-based methods and the statistical efficiency of parametric methods for small sample sizes. It is a computationally feasible framework that addresses all three issues simultaneously. We show that our method controls false positives at desired significance levels. An additional advantage is its flexibility in modeling various types of single-cell data. The usefulness of our method is further demonstrated in two studies: a single-cell proteomics study showing the much improved efficiency of our method than the pseudobulk version of the Wilcoxon rank-sum test, and a single-cell RNA study, where our method identified genes that would be missed without

adjusting for covariates.

10:40 AM–11:00 AM Speaker: Rong Zablock, University of California, San Diego

Novel statistical inference tools for longitudinal changes in accelerometer-measured physical activity
Author(s): Rong W. Zablocki, University of California at San Diego; Andrea Z. LaCroix, University of California at San Diego; Loki Natarajan, University of California at San Diego; Lindsay Dillon, University of California at San Diego; Jingjing Zou, University of California at San Diego

The wide usage of wearable accelerometer-based activity trackers in recent years has provided a unique opportunity for in-depth research on physical activity (PA) and its relationship with health outcomes and interventions. Past analysis of activity tracker data relies heavily on aggregating minute-level PA records into day-level summary statistics, in which important information of diurnal PA patterns is lost. In this talk I will introduce our recent development of a novel functional data analysis approach based on theory of Riemann manifolds for longitudinal changes in PA patterns. We further examine in this work the statistical inference tools tailored for the detecting, testing, and clustering of longitudinal changes in PA diurnal patterns. With the proposed approaches we conduct comprehensive analyses on data from the Women's Health Initiative (WHI).

11:00 AM–11:20 AM Speaker: Esra Kurum, University of California, Riverside

Spatiotemporal multilevel joint modeling of generalized longitudinal and survival outcomes in end-stage kidney disease
Author(s): Esra Kurum, University of California, Riverside

Individuals with end-stage kidney disease (ESKD) on dialysis experience high mortality and excessive burden of hospitalizations over time relative to comparable Medicare patient cohorts without kidney failure. A key interest in this population is to understand the time-dynamic effects of multilevel risk factors that contribute to the correlated outcomes of longitudinal hospitalization and mortality. We utilize multilevel data from the United States Renal Data System (USRDS), where repeated measurements/hospitalizations over time are nested in patients and patients are nested within (health service) regions across the U.S. We develop a novel spatiotemporal multilevel joint model (STM-JM) that accounts for the aforementioned hierarchical structure of the data while considering the spatiotemporal variations in both outcomes across regions. The proposed STM-JM includes time-varying effects of multilevel (patient- and region-level) risk factors on the correlated outcomes and incorporates spatial correlations across the spatial regions via a multivariate conditional autoregressive correlation structure. Efficient estimation and inference are performed via a Bayesian framework. An application of the proposed method to the USRDS data highlights significant time-varying effects of risk factors on hospitalization and mortality and identifies specific time periods on dialysis and spatial locations across the U.S. with elevated hospitalization and mortality risks.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S38: PRACTICAL CONSIDERATIONS IN USING WIN STATISTICS

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Gold (Lower Level)

Organizer: Yu Cheng, University of Pittsburgh
Chair: Yu Cheng, University of Pittsburgh

10:00 AM–10:20 AM Speaker: Roland Matsouaka, Duke University

Trial design with win ratio or win odds based on hierarchical endpoints

Author(s): Roland Matsouaka, Duke University; Huiman Barnhardt, Duke University; Yuliya Lokhnygina, Duke University; Frank Rockhold, Duke University

Win statistics, such as win ratio and win odds, have become a popular approach to analysis of hierarchical endpoints in clinical studies. While several sample size or power calculation formulas are available for the design of randomized trials using these methods, these formulas require investigators to specify clinically significant and meaningful magnitude of win ratio or win odds, as well as the expected probability of ties. In practice, these quantities are difficult to come up with based on prior published literature. In this presentation, we show that the win ratio for the hierarchical endpoints is a weighted average of marginal win ratios (with similar expression for win odds), under the assumption of independence of the individual endpoints. We also provide the expression for the probability of ties. With only the need to specify clinically significant marginal win ratios (or win odds), these formulas provide a simple way to come up with clinically significant win ratio (or win odds) and probability of ties that are defendable. As a result, formula-based power and sample size calculations can be easily obtained for trial design without the need of conducting complex simulation studies. Our extensive simulation studies show that powers calculated with the formulas under independence assumption are similar to the simulated powers for any type of positively correlated hierarchical endpoints. Our approach provides researchers an easy tool for trial design as well as providing insights on relative contribution of marginal win ratio (or win odds) on overall win ratio (or win odd), and the impact of adding additional endpoint to the hierarchy. Cardiovascular trials are used to illustrate our approaches in trial design with mixed types of hierarchical endpoints.

10:20 AM–10:20 AM Speaker: Dali Zhou, US Food and Drug Administration

Discussion on the use of win ratio approach in confirmatory analyses
Author(s): Dali Zhou, US Food and Drug Administration

Discussions of the advantages and challenges on use of win ratio approach in confirmatory analyses from regulatory perspective.

10:40 AM–11:00 AM Speaker: Bang Wang, Vertex Pharmaceuticals

Generalized win-odds regression models for composite endpoints
Author(s): Bang Wang, Zi Wang, Yu Cheng

The time-to-first-event analysis is often used for studies involving multiple event times, where each component is treated equally, regardless of their clinical importance. Alternative summaries such as Win Ratio, Net Benefit, and Win Odds (WO) have drawn attention lately because

they can handle different types of outcomes and allow for a hierarchical ordering in component outcomes. In this paper, we focus on WO and propose proportional WO regression models to evaluate the treatment effect on multiple outcomes while controlling for other risk factors. The models are easily interpretable as a standard logistic regression model. However, the proposed WO regression is more advanced; multiple outcomes of different types can be modeled together, and the estimating equation is constructed based on all possible and potentially dependent pairings of a treated individual with a control one under the functional response modeling framework. In addition, informative ties are carefully distinguished from those inconclusive comparisons due to censoring, and the latter is handled via the inverse probability of censoring weighting method. We establish the asymptotic properties of the estimated regression coefficients using the U-statistic theory and demonstrate the finite sample performance through numerical studies.

11:00 AM–11:20 AM Discussant: David Oakes, University of Rochester

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S39: METHODS AND APPLICATIONS FOR ANALYZING MULTIPLE ENDPOINTS OR MULTIVARIATE RESPONSES

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Melody (Lobby Level)

Organizer: Qingcong Yuan, Sanofi
Chair: Hong Wang, University of Pittsburgh

10:00 AM–10:20 AM Speaker: Kexuan Li, Bristol Myers Squibb

Multivariate rank-based analysis of multiple endpoints in clinical trials: A global test approach

Author(s): Kexuan Li, Lingli Yang, Shaofei Zhao, Susie Sinks, Luan Lin, Peng Sun

Clinical trials often involve the assessment of multiple endpoints to comprehensively evaluate the efficacy and safety of interventions. In the work, we consider a global nonparametric testing procedure based on multivariate rank for the analysis of multiple endpoints in clinical trials. Unlike other existing approaches that rely on pairwise comparisons for each individual endpoint, the proposed method directly incorporates the multivariate ranks of the observations. By considering the joint ranking of all endpoints, the proposed approach provides robustness against diverse data distributions and censoring mechanisms commonly encountered in clinical trials. Through extensive simulations, we demonstrate the superior performance of the multivariate rank-based approach in controlling type I error and achieving higher power compared to existing rank-based methods. The simulations illustrate the advantages of leveraging multivariate ranks and highlight the robustness of the approach in various settings. The proposed method offers an effective tool for the analysis of multiple endpoints in clinical trials, enhancing the reliability and efficiency of outcome evaluations.

10:20 AM–10:20 AM Speaker: Xiaodong Luo, Sanofi

Multiplicity control in adaptive and nonadaptive sequential design with multiple endpoints and/or arms

Author(s): Xiaodong Luo, Sanofi

Multiplicity control in sequential design with multiple endpoints have constantly posted challenges to clinical trials, such problem becomes more difficult when adaptive designs of selecting treatments and/or populations are also considered. In this talk, I will propose a simple framework to tackle these challenges and demonstrate that some handy solutions naturally arise from this framework.

10:40 AM–11:00 AM Speaker: Xiaoli Kong, Wayne State University

Multivariate rank-based nonparametric measures for testing independence

Author(s): Xiaoli Kong, Wayne State University; Qingcong Yuan, Sanofi US; Chenlu Ke, Virginia Commonwealth University

This presentation introduces a class of multivariate rank-based measures for testing independence, which utilize the expected difference between conditional and marginal characteristic functions. The tests are computationally efficient and remain well-defined under minimal assumptions. We establish the asymptotic distribution for these test statistics and demonstrate

their effectiveness through simulation studies. Additionally, we illustrate the practical utility of the proposed test through real data applications.

11:00 AM–11:20 AM Speaker: Chenlu Ke, Virginia Commonwealth University

Reimaging semi-competing risks data analysis: Enhancing variable selection with preliminary sufficient dimension reduction

Author(s): Chenlu Ke, Virginia Commonwealth University

We introduce a new framework with an efficient algorithm for feature screening in the challenging context of ultrahigh dimensional semi-competing risks data. Specifically, our two-stage procedure initially employs a dual screening mechanism to select a coarse set of features that are potentially relevant to both terminal and nonterminal endpoints. This leads to the estimation of the augmented central subspace, pivotal for both endpoints and censoring, based on the selected features. In the second stage, refined sets of important features for the nonterminal and terminal events, respectively, are further identified using an inverse probability-of-censoring weighted filter, where the central subspace estimator is used to obtain the weights adjusting for censoring. The proposed framework is model-free and it does not require independent censoring. Asymptotic properties are established under minor assumptions. We demonstrate the promising performance of the proposed method through simulations and gene expression data analysis.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S40: NEW FRONTIERS IN MACHINE LEARNING AND STATISTICAL LEARNING

Tuesday, June 18, 2024
10:00 AM–11:30 AM, Green Room (Lobby Level)

Organizer: Yichuan Zhao, Georgia State University
Chair: Song Yang, National Heart, Lung, and Blood Institute (NHLBI), NIH

10:00 AM–10:20 AM Speaker: Xiaojing Wang, University of Connecticut

Bayesian clustering of subpopulations in neural spiking activity

Author(s): Ganchao Wei, Duke University; Ian Stevenson, University of Connecticut; Xiaojing Wang, University of Connecticut

With the advanced capability to record the spiking activity of many hundreds of neurons simultaneously, new statistical methods are needed to understand the structure of this large-scale neural population activity. Although previous work has tried to summarize neural activity within and between known populations by extracting low-dimensional latent factors, in many cases what determines a unique population may be unclear. Neurons may differ in anatomical location, cell types or response properties. To identify populations directly related to neural activity, we propose a flexible clustering method based on a mixture of dynamic Poisson factor analyzers (mixDPFA) model, with both the number of clusters and dimension of latent factors for each cluster are unknown, which makes the analysis of the mixDPFA model very challenging. We develop a Markov chain Monte Carlo (MCMC) algorithm that makes the computation feasible and can efficiently sample from the posterior distribution to make inferences. Validating our proposed MCMC algorithm through simulations, we find that it can accurately recover the unknown parameters, the dimensions of latent factors and the true clustering in the model. These results are insensitive to the initial cluster assignments. We then apply the proposed mixDPFA model to multi-region experimental recordings, where we find that the proposed method can identify novel, reliable clusters of neurons based on their activity, and may, thus, be a useful tool for neural data analysis.

10:20 AM–10:20 AM Speaker: Yiyuan She, Florida State University

Enhancing federated learning with range penalization

Author(s): Yiyuan She, Florida State University

In this talk, we introduce a new method for federated learning that tackles data heterogeneity across clients. Our approach not only identifies shared features but also adaptively forms clusters for personalized features at extreme values of their coefficients, all within a reduced range. This technique significantly improves data transmission efficiency and optimizes storage within federated learning frameworks. We conduct nonasymptotic analyses to evaluate the statistical accuracy of our new regularization approach compared to existing methods. We also develop a fast optimization algorithm that utilizes local strong convexity and smoothness to accelerate the learning process. Experiments confirm the effectiveness and efficiency of our proposed method.

10:40 AM–11:00 AM Speaker: Lily Wang, George Mason University

Distributed heterogeneity learning: From spatial to complex data analysis

Author(s): Shan Yu, University of Virginia; Guannan Wang, William & Mary; Lily Wang, George Mason University

Heterogeneity learning is a fundamental challenge spanning various scientific domains, from social, economic, and environmental studies to the broader landscape of complex data analysis. To address this challenge, spatially varying coefficient models have emerged as potent tools for tackling spatial regression heterogeneity. This presentation introduces a class of generalized partially linear spatially varying coefficient models that enable the inclusion of both constant and spatially varying covariate effects while balancing flexibility and parsimony. In addition, to address the challenge of extraordinarily large and complex datasets collected from modern technologies, we propose a novel distributed heterogeneity learning (DHL) method based on multivariate spline smoothing over a triangulation of the domain. The DHL algorithm has a simple, scalable, and communication-efficient implementation scheme that can almost achieve linear speedup. The DHL framework is theoretically supported by demonstrating the asymptotic normality of DHL linear estimators and DHL spline estimators' convergence rate equivalent to that of global spline estimators obtained from the entire dataset. We further expand the scope of DHL to a wider range of varying coefficient models, broadening its applicability to complex data analysis domains such as spatiotemporal data, functional data, and point cloud learning. The efficacy of the extended DHL is evaluated through comprehensive simulation studies and real-world applications.

11:00 AM–11:20 AM Speaker: Samuel Wu, University of Florida

Statistical analyses for differentially-private matrix masking data

Author(s): Linh Nghiem, University of Sydney; Aidong Ding, Northeastern University; Samuel Wu, University of Florida

A significant challenge in big data is finding a balance between preserving individuals' privacy and maintaining data utility for statistical analysis. While differential privacy (DP) quantifies privacy protection, statistical inference methods for privacy-protected data remain underexplored. In this paper, we introduce statistical methods for analyzing DP-guaranteed datasets based on a combination of triple matrix masking and noise addition. The first analysis focuses on the linear regression model, emphasizing the connection and differences between privacy protection and traditional measurement error settings. The second analysis is on a 2x2 contingency table, where we propose valid statistical inference procedures for the difference in two proportions. Our theoretical findings reveal a trade-off between privacy protection and statistical precision. Finally, we apply these methods to a dataset concerning hypertension prevalence in the United States, comparing insights derived from the original data and the privacy-protected release.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S41: DEEP LEARNING FOUNDATION AND APPLICATIONS

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Symphony 1 (Lobby Level)

Organizer: Xiao Wang, Purdue University
Chair: Runpeng Dai, University of North Carolina at Chapel Hill

1:00 PM–1:20 PM Speaker: Faming Liang, Purdue University

On causal inference with deep neural networks

Author(s): Yaxin Fang, Purdue University; Faming Liang, Purdue University

With the advancement of data science, the collection of increasingly complex datasets has become commonplace. In such datasets, the data dimension can be extremely high, and the underlying data generation process can be unknown and highly nonlinear. As a result, the task of making causal inference with high-dimensional complex data has become a fundamental problem in many disciplines, such as medicine, econometrics, and social science. However, the existing methods for causal inference are frequently developed under the assumption that the data dimension is low or that the underlying data generation process is linear or approximately linear. To address these challenges, we propose a novel causal inference approach for dealing with high-dimensional complex data. The proposed approach is based on deep learning techniques, including sparse deep learning theory and stochastic neural networks, that have been developed in recent literature. By using these techniques, the proposed approach can address both the high dimensionality and unknown data generation process in a coherent way. Furthermore, the proposed approach can also be used when missing values are present in the datasets. Extensive numerical studies indicate that the proposed approach outperforms existing ones.

1:20 PM–1:20 PM Speaker: Rui Feng, University of Pennsylvania

Structure guided transformer clustering

Author(s): Rui Feng, University of Pennsylvania

Deep learning clustering methods has become increasingly popular in high-dimensional data. They learn low-dimensional feature representations from complex data and then apply classic clustering methods on the derived representations, which offers flexibility for diverse data types and has demonstrated advantages over traditional clustering methods. However, current deep learning clustering methods treat all input variables equally, potentially leading to an imbalance in similarity matrices and producing clustering results that lack meaningful interpretation. To address this issue and reduce redundancy among predictors, we proposed a novel clustering method based on transformer models, using the inherent structural correlations among predictors. We evaluated the performance of our proposed method through simulations and comparisons with existing approaches. We applied our method to TCGA AML patients to identify distinct clusters. The clusters obtained using our approach showed stronger correlations with survival outcomes of the patients compared to other existing methods.

1:40 PM–2:00 PM Speaker: Yun Yang, University of Illinois Urbana-Champaign

Adaptivity of diffusion models to manifold structures
Author(s): Yun Yang, University of Illinois Urbana-Champaign; Rong Tang, Hong Kong University of Science and Technology

Empirical studies have demonstrated the effectiveness of (score-based) diffusion models in generating high-dimensional data, such as texts and images, which typically exhibit a low-dimensional manifold nature. These empirical successes raise the theoretical question of whether score-based diffusion models can optimally adapt to low-dimensional manifold structures. While recent work has validated the minimax optimality of diffusion models when the target distribution admits a smooth density with respect to the Lebesgue measure of the ambient data space, these findings do not fully account for the ability of diffusion models in avoiding the the curse of dimensionality when estimating high-dimensional distributions. This work considers two common classes of diffusion models: Langevin diffusion and forward-backward diffusion. We show that both models can adapt to the intrinsic manifold structure by showing that the convergence rate of the inducing distribution estimator depends only on the intrinsic dimension of the data. Moreover, our considered estimator does not require knowing or explicitly estimating the manifold. We also demonstrate that the forward-backward diffusion can achieve the minimax optimal rate under the Wasserstein metric when the target distribution possesses a smooth density with respect to the volume measure of the low-dimensional manifold.

2:00 PM–2:20 PM Speaker: Runpeng Dai, University of North Carolina at Chapel Hill

Causal deepsets for off-policy evaluation under spatial or spatio-temporal interferences
Author(s): Runpeng Dai, University of North Carolina at Chapel Hill

Off-policy evaluation (OPE) is widely applied in sectors such as pharmaceuticals and e-commerce to evaluate the efficacy of novel products or policies from offline datasets. This paper introduces a causal deepset framework that relaxes several key structural assumptions, primarily the mean-field assumption, prevalent in existing OPE methodologies that handle spatio-temporal interference. These traditional assumptions frequently prove inadequate in real-world settings, thereby restricting the capability of current OPE methods to effectively address complex interference effects. In response, we advocate for the implementation of the permutation invariance (PI) assumption. This innovative approach enables the data-driven, adaptive learning of the mean-field function, offering a more flexible estimation method beyond conventional averaging. Furthermore, we present novel algorithms that incorporate the PI assumption into OPE and thoroughly examine their theoretical foundations. Our numerical analyses demonstrate that this novel approach yields significantly more precise estimations than existing baseline algorithms, thereby substantially improving the practical applicability and effectiveness of OPE methodologies.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S42: FRONTIERS OF MEDICAL DECISION-MAKING IN THE DATA-DRIVEN ERA

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Blackbird A (Mezzanine Level)

Organizer: Yingqi Zhao, Fred Hutchinson Cancer Center
Chair: Bo Zhang, Fred Hutchinson Cancer Center

1:00 PM–1:20 PM Speaker: Xingche Guo, Columbia University

Characterizing human reward-based decision-making behavior with reinforcement learning models

Author(s): Xingche Guo, Columbia University; Donglin Zeng, Michigan University; Yuanjia Wang, Columbia University

Major depressive disorder (MDD) is one of the leading causes of disability-adjusted life years. Emerging evidence indicates that reward processing abnormalities may serve as a behavioral marker for MDD. To measure reward processing, patients perform computer-based behavioral tasks that involve making choices based on different stimuli, such as rewards and penalties. Reinforcement learning (RL) are widely used to characterize how patients make decisions in reward-based behavioral tasks. To account for the nonlinearity of the decision process, we propose a semiparametric RL (Semi-RL) approach that models the RL parameters with nonparametric functions. Between-subject heterogeneity is considered by incorporating random effects. We provide a computationally efficient solution to address the challenges posed by nonparametric functions and random effects, along with theoretical results demonstrating the consistency and asymptotic normality of the parameters of interest. In the real data analysis, we show that individuals with MDD exhibit lower reward sensitivity compared to healthy subjects, and the reward sensitivity has a nonlinear form with a floor and ceiling effect. This finding suggests a switching of decision-making processes between multiple learning strategies. In my recent work, I propose a RL with hidden Markov models (RL-HMM) framework, enabling learning strategy switching between two distinct strategies: RL model or random choices. The computational algorithm via EM algorithm is briefly introduced. Utilizing the RL-HMM, we demonstrated that individuals with MDD face greater difficulty concentrating during tasks compared to the healthy subjects. Finally, a brief overview of brain-behavior associations is provided, exploring potentials for integrating behavioral data with neuroimaging modalities such as EEG and fMRI.

1:20 PM–1:20 PM Speaker: Shu Yang, North Carolina State University

Multiply robust off-policy evaluation and learning under truncation by death

Author(s): Jianing Chu, North Carolina State University; Shu Yang, North Carolina State University; Wenbin Lu, North Carolina State University

Typical off-policy evaluation (OPE) and off-policy learning (OPL) are not well-defined problems under "truncation by death", where the outcome of interest is not defined after some events, such as death. The standard OPE no longer yields consistent estimators, and the standard OPL results in suboptimal policies. In this paper, we formulate OPE and OPL using principal stratification under "truncation by death." We propose a survivor value function for a subpopulation whose outcomes are always defined regardless of treatment conditions. We establish a novel

identification strategy under principal ignorability, and derive the semiparametric efficiency bound of an OPE estimator. Then, we propose multiply robust estimators for OPE and OPL. We show that the proposed estimators are consistent and asymptotically normal even with flexible semi/nonparametric models for nuisance functions approximation. Moreover, under mild rate conditions of nuisance functions approximation, the estimators achieve the semiparametric efficiency bound. Finally, we conduct experiments to demonstrate the empirical performance of the proposed estimators.

1:40 PM–2:00 PM Speaker: Nilanjana Laha, Texas A&M University

Optimal dynamic treatment regimes via smooth surrogates

Author(s): Nilson Chapagain, Texas A&M University; Aarón Sonabend W, Google Research; Nilanjana Laha, Texas A&M University

During the treatment of chronic diseases, such as cancer, sepsis, and diabetes, patients may receive treatments multiple times. Our aim is to learn the best sequence of treatments, also called a treatment policy or dynamic treatment regimes, using available patient data such as electronic health record data. Although the DTR learning problem is an offline reinforcement learning (RL) problem, most standard offline RL methods are unsuitable for the DTR setting due to their inability to leverage the whole patient history. A recent direction of DTR research has shown that efficient DTR learning is possible via direct policy search, but computationally feasible algorithms for the latter are currently available only in binary treatment cases. This project aims to develop a direct policy search framework for DTR problems for general cases. The proposed methods are model free, and reduces to non-convex but smooth optimization problems. I will introduce the proposed method, show some experimental and theoretical guarantees, and discuss some open questions.

2:00 PM–2:20 PM Speaker: Yingqi Zhao, Fred Hutchinson Cancer Center

Improve fairness of Neyman-Pearson classifiers under data shift

Author(s): Jiaming Qiu, Fred Hutchinson Cancer Center; Yingqi Zhao, Fred Hutchinson Cancer Center; Yingye Zheng, Fred Hutchinson Cancer Center

Clinical decision rules incorporating novel biomarkers to balance benefits and risks are desired in early cancer detection. However, the practical application of these decision rules encounters challenges stemming from discrepancies in data distribution between the training population and the intended application population. The potential impact can be disproportional particularly if under-represented groups undergo a data shift from that of the training data. We highlight a semi-parametric model-based approach to adapt biomarker-assisted rules for differing populations while controlling true or false positive rates targeting specific clinical applications. We propose a calibrating strategy to utilize a small set of additional unlabeled testing data, with minimal auxiliary information alongside the labeled training data, to improve the fairness of the learned decision rule. This approach consistently tailors decision rules for the target population, as demonstrated through both theoretical studies and simulations. We illustrate the approaches with an example of a prostate cancer study.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S43: UNDERSTANDING THE HETEROGENEITY OF GENETICS EFFECTS, FROM SINGLE CELLS TO BIOBANK DATA.

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Lyric (Lobby Level)

Organizer: Wei Sun, Fred Hutchinson Cancer Center
Chair: Tianying Wang, Colorado State University

1:00 PM–1:20 PM Speaker: Lulu Shang, University of Texas MD Anderson Cancer Center

Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics

Author(s): Lulu Shang, University of Texas MD Anderson Cancer Center; Peijun Wu, University of Michigan; Xiang Zhou, University of Michigan

An essential task in spatial transcriptomics involves identifying genes with spatial expression patterns, known as spatially variable genes (SVGs). Importantly, a subset of SVGs displays diverse spatial expression patterns within a given cell type, thus representing key transcriptomic signatures underlying cellular heterogeneity. Here, we present Celina, a statistical method for systematically detecting this subset of cell type-specific SVGs (ct-SVGs). Celina utilizes a spatially varying coefficient model to accurately capture each gene's spatial expression pattern in relation to the distribution of cell types across tissue locations, ensuring effective type I error control and high statistical power. We evaluated the performance of Celina through comprehensive simulations and applications to five real datasets, where we also adapted and examined existing methods originated from other analytic settings to detect ct-SVGs. Celina proves powerful compared to these ad hoc method adaptations in single cell resolution spatial transcriptomics and stands as the only effective solution for spot resolution spatial transcriptomics. In the real data applications, Celina uncovers ct-SVGs associated with tumor progression and patient survival in lung cancer, identifies metagenes with unique spatial patterns linked to cell proliferation and immune response in kidney cancer, and detects genes preferentially expressed near amyloid-β plaques in an Alzheimer's model. The ct-SVGs detected by Celina also enable novel biologically informed downstream analyses, unveiling functional cellular heterogeneity at an unprecedented scale.

1:20 PM–1:20 PM Speaker: Fan Wang, Columbia University

Computationally efficient whole-genome quantile inference with related samples

Author(s): Fan Wang, Columbia University; Tianying Wang, University of Colorado; Chen Wang, Columbia University; Iuliana Ionita-Laza, Columbia University

Quantile regression is an alternative to linear regression for quantitative phenotypes that allows detection of heterogeneous genotype-phenotype associations. We propose here novel and computationally efficient extensions of quantile regression to performing GWAS with related samples. We consider two novel tests, one based on weighted quantile rank score test, and a second one based on the offset method as previously used in the context of whole-genome regression method Regenie. We show good control of type-1 error in simulations and present results of applications to UK Biobank.

1:40 PM–2:00 PM Speaker: Wenmin Zhang, McGill University

Accounting for genetic effect heterogeneity in fine-mapping and improving power to detect gene-environment interactions with SharePro

Author(s): Wenmin Zhang, Montreal Heart Institute / McGill University; Rob Sladek, McGill University; Yue Li, McGill University; Hamed Najafabadi, McGill University; Josée Dupuis, McGill University

Characterizing genetic effect heterogeneity across subpopulations with different environmental exposures is useful for identifying exposure-specific pathways, understanding biological mechanisms underlying disease heterogeneity and further pinpointing modifiable risk factors for disease prevention and management. Classical gene-by-environment interaction (GxE) analysis can be used to characterize genetic effect heterogeneity. However, it can have a high multiple testing burden in the context of genome-wide association studies (GWAS) and requires a large sample size to achieve sufficient power. We adapt a colocalization method, SharePro, to account for effect heterogeneity in fine-mapping and subsequently improve power for GxE analysis. Through joint fine-mapping of exposure stratified GWAS summary statistics, SharePro can greatly reduce multiple testing burden in GxE analysis. Through extensive simulation studies, we demonstrated that accounting for effect heterogeneity can improve power for fine-mapping. With efficient joint fine-mapping of exposure stratified GWAS summary statistics, SharePro alleviated multiple testing burden in GxE analysis and demonstrated improved power with a well-controlled false discovery rate. Through analyses of smoking status stratified GWAS summary statistics, we identified genetic effects on lung function modulated by smoking status mapped to the genes CHRNA3, ADAM19 and UBR1. Additionally, using sex stratified GWAS summary statistics, we characterized sex differentiated genetic effects on fat distribution and provided biologically plausible candidates for functional follow-up studies. In summary, we have developed an analytical framework to account for effect heterogeneity in fine-mapping and subsequently improve power for GxE analysis. The SharePro software for GxE analysis is openly available at .

2:00 PM–2:20 PM Speaker: Guanghao Qi, University of Washington

Single-cell allele-specific expression analysis reveals dynamic and cell-type-specific regulatory effects

Author(s): Guanghao Qi, University of Washington; Benjamin J. Strober, Harvard University; Joshua M. Popp, Johns Hopkins University; Rebecca Keener, Johns Hopkins University; Hongkai Ji, Johns Hopkins University; Alexis Battle, Johns Hopkins University

Differential allele-specific expression (ASE) is a powerful tool to study context-specific cis-regulation of gene expression. Such effects can reflect the interaction between genetic or epigenetic factors and a measured context or condition. Single-cell RNA sequencing (scRNA-seq) allows the measurement of ASE at individual-cell resolution, but there is a lack of statistical methods to analyze such data. We present Differential Allelic Expression using Single-Cell data (DAESC), a powerful method for differential ASE analysis using scRNA-seq from multiple individuals, with statistical behavior confirmed through simulation. DAESC accounts for non-independence between cells from the same individual and incorporates implicit haplotype phasing. Application to data from 105 induced pluripotent stem cell (iPSC) lines identifies 657

genes dynamically regulated during endoderm differentiation, with enrichment for changes in chromatin state. Application to a type-2 diabetes dataset identifies several differentially regulated genes between patients and controls in pancreatic endocrine cells. DAESC is a powerful method for single-cell ASE analysis and can uncover novel insights on gene regulation.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S44: INSIGHTS IN CLINICAL TRIALS / DRUG DEVELOPMENT

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Ocean Way (Mezzanine Level)

Organizer: Xiyuan Gao, AbbVie Inc.
Chair: Yan Zhuang, Indiana University

1:00 PM–1:20 PM Speaker: Yong Lin, Eli Lilly and Company

Challenges, reflections, and insights in early-phase clinical trial development

Author(s): Yong Lin, Eli Lilly and Company

Early-phase clinical trials play a pivotal role in drug development, offering crucial insights into safety, dosing, and pharmacokinetics. These trials occur during the initial stages of testing and typically involve healthy volunteers or a small cohort of patients. However, the involvement of statisticians is critical even before patient enrollment. They contribute to toxicology study planning and drug material preparation, which ultimately informs the duration and detailed design of single-dose escalation and multiple-dose escalation strategies in Phase I studies. In many cases, the development strategy is guided by a delicate balance between expediting progress to Phase II and maximizing the depth of information gained from Phase I. In this discussion, we explore examples drawn from early-phase development, including the duration of Phase I multiple ascending dose studies and the challenges of dose finding as trials progress from Phase I to Phase II. These examples underscore the complexities inherent in early-phase drug development, encouraging thoughtful and ultimately hope to inspire innovative solutions.

1:20 PM–1:20 PM Speaker: Naitee Ting, Boehringer Ingelheim Pharmaceuticals, Inc.

Tables in a clinical study report—quality or quantity?

Author(s): Naitee Ting, Boehringer Ingelheim Pharmaceuticals, Inc.

A recent phenomenon is that there are too many tables in a clinical study report (CSR). The main reason for this to happen is that the clinical statisticians in pharmaceutical industry (or in contract research organizations) misunderstand their role/responsibility in a clinical team. In fact the two most important contributions a statistician can make to the team or to the company is study design and statistical consulting. If a statistician does a good job in these contributions, it would only be natural that certain tables are key tables to help with decision making, some tables provide information to help better understand various properties of the drug. All other tables are not necessary. For any statistical analysis plan (SAP), the concern is not number of tables, but each proposed table has to support the understanding and decision for the study drug. In this presentation, we will emphasize the values of statisticians in clinical trials, dig into the reasons of excessive tables planned in SAP and generated in CSR, use two hypothetical dialog examples to illustrate the statistical thinking on quality versus quantity, and finally conclude this presentation with a brief remark.

1:40 PM–2:00 PM Speaker: Li Wang, AbbVie Inc.

Artificial intelligence demystified and how can it help clinical development
Author(s): Li Wang, AbbVie Inc.

ML and AI are widely used in technology industry now and are finding their ways into drug development in pharmaceutical industry especially in manufacturing and discovery. In clinical development, how and where to appropriately apply ML/AI and what value it can bring to the business remain big questions without clear answers to researchers. For traditional clinical statisticians, it is not only a big challenge but also an exciting opportunity. AbbVie's statistical innovation group closely collaborated with different development functions to leverage multiple data sources and deep learning methodologies to help make smarter decisions. Experiences and some use cases will be shared.

2:00 PM–2:20 PM Speaker: Yan Zhuang, Indiana University

Enhancing decentralized clinical trial management through blockchain technology
Author(s): Yan Zhuang, Indiana University

In the evolving landscape of clinical trials, decentralized clinical trials (DCTs) are emerging as a revolutionary model, offering enhanced patient accessibility and improved data quality. However, they introduce unique management challenges. This presentation delves into the integration of blockchain technology, renowned for its robust transparency, traceability, immutability, and security in the financial sector, as a solution to these DCT-specific challenges. We propose a blockchain-based system tailored for DCT management, encompassing essential features such as a universally accessible trial master file system, streamlined patient recruitment and enrollment, secure electronic data capture, and comprehensive data analytics. This system also efficiently manages financial transactions within clinical trials. The successful implementation and empirical testing of our blockchain-based solution highlight its capacity to address the complexities of DCT management effectively. This advancement marks a significant stride in utilizing technology to enhance clinical trial processes, promoting more efficient, secure, and patient-centric trials in the decentralized era.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S45: COVARIATE-ADAPTIVE RANDOMIZATION: METHODS AND APPLICATIONS

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Platinum (Lower Level)

Organizer: Xiaotian Chen, AbbVie Inc.
Chair: Yu Deng, AbbVie Inc.

1:00 PM–1:20 PM Speaker: Hongjian Zhu, AbbVie Inc.

Seamless phase II/III clinical trials with covariate adaptive randomization

Author(s): Wei Ma, China Renmin University; Mengxi Wang, Univeristy of Texas Health Science Center; Hongjian Zhu, AbbVie Inc.

There is an urgent need to evaluate new therapies in a time-sensitive and cost-effective manner. We propose the adaptive seamless phase II/III clinical trials with covariate adaptive randomization (CAR) to satisfy this need. CAR is one of the most popular designs in randomized controlled trials, enhancing covariance balance and ensuring valid treatment comparisons. However, it has several challenges: (1) the type I error rate of the commonly used Student's t-test following CAR can be inflated because of the seamless trials, but can also be decreased using CAR; (2) the complicated allocation mechanism induced by CAR causes extra difficulties to derive the asymptotic properties of a test procedure; and (3) previous theoretical studies of seamless trials rely mainly on the assumption of complete randomization, a procedure rarely used in real trials. We establish a theoretical foundation for adaptive seamless phase II/III trials with CAR. We also propose an approach that is easy to implement in order to control the type I error rate and improve the power when using Student's t-test. This important step will promote the application of this procedure.

1:20 PM–1:20 PM Speaker: Fengyu Zhao, The George Washington University

Inference of G-computation methods under covariate adaptive randomization

Author(s): Fengyu Zhao, The George Washington University

In clinical research, there is an increasing interest in employing G-computation methods for estimating treatment effects, due to their superior statistical power and robustness to model misspecification. Traditionally, the inferential properties of G-computation have been concentrated within the domain of complete randomization. However, in the context of covariate adaptive randomization (CAR), the effectiveness of G-computation methods is not assured due to their usage on covariate information for patient allocation. This presentation introduces the use of G-computation through a logistic regression model, also known as marginal standardization, to evaluate relative risk. We highlight that while traditional G-computation tests are valid in completely randomized settings, their validity may be compromised in CAR situations. To address this, we propose a variance adjustment technique that aims to increase statistical power while maintaining accurate type I error rates in CAR contexts. Numerical studies support our theoretical properties and showcase the efficacy of our proposed adjustment method.

1:40 PM–2:00 PM Speaker: Jialu Wang, Vertex Pharmaceuticals

Covariate-adaptive design in network data

Author(s): Jialu Wang, Vertex Pharmaceuticals; Ping Li, LinkedIn Corporation; Feifang Hu, George Washington University

People linked together through a network often tend to have similar behaviors. This phenomenon is usually known as network interaction. Their covariates are often correlated with their outcomes as well. Therefore, one should incorporate both the covariates and the network information in a carefully designed randomization to improve the estimation of the average treatment effect (ATE) in network data. In this talk, we introduce a new adaptive design to balance both the network and the covariates. We show that the imbalance measures with respect to the covariates and the network are Op(1). We also demonstrate the relationships between the improved balances and the increased efficiency in terms of the mean square error (MSE). Numerical studies demonstrate the advanced performance of the proposed design regarding the greater comparability of the treatment groups and the reduction of MSE for estimating the ATE.

2:00 PM–2:20 PM Speaker: Xiaotian Chen, AbbVie Inc.

Sequential monitoring of the average treatment effect in randomized controlled trials with covariate-adaptive randomization

Author(s): Xiaotian Chen, AbbVie; Jun Yu, AbbVie; Hongjian Zhu, AbbVie; Li Wang, AbbVie

For randomized clinical trials in drug development and clinical research, minimizing the imbalance of treatment allocation among key baseline characteristics is crucial due to its impact on the precision of estimation of the treatment effect and related inferences. Covariate-adaptive randomization (CAR) methods have been used to achieve balance and appropriate analysis approaches have been developed to address the confounding problem associated with the CAR mechanism. The nonparametric inference method recently proposed by Bugni and colleagues provided adjustment to the usual hypothesis tests to address their conservativeness. We extend such approach in the sequential monitoring framework with interim analyses, motivated by the ethical and economic advantages of such design with the opportunities of early stopping. We provide theoretical results and perform numerical studies to demonstrate the improvement on statistical efficiency including the power to detect meaningful treatment effect. A real example of study redesign will also be discussed.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S46: ADVANCED METHODS FOR FEDERATED AND TRANSFER LEARNING WITH APPLICATION IN REAL-WORLD DATA

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Sound Emporium A/B (Mezzanine Level)

Organizer: Xiaokang Liu, University of Missouri
Chair: Xiaokang Liu, University of Missouri

1:00 PM–1:20 PM Speaker: Rui Duan, Harvard University

Collaborative statistical learning with weakly aligned source

Author(s): Rui Duan, Harvard University

The growing potential for multi-institutional collaborative research and data integration is unlocking new possibilities in statistical learning and inference, yet it comes with its own set of challenges. Our talk highlights the advancements in collaborative statistical inference, addressing key issues such as data heterogeneity, model misspecifications, and the complexities of data sharing. We will introduce several methods we designed and illustrate the considerations from statistical efficiency, robustness against diverse data sources and communication constraints. Finally, we will present a comprehensive examination of the proposed approaches through theoretical analysis, numerical experiments, and the application to real-world datasets.

1:20 PM–1:20 PM Speaker: Yuekai Sun, University of Michigan

Equality and equity in performative prediction

Author(s): Seamus Somerstep, University of Michigan; Ya'acov Ritov, University of Michigan; Yuekai Sun, University of Michigan

In many prediction problems, the predictive model affects the distribution of the prediction target. This phenomenon is known as performativity, and it is often caused by the behavior of individuals with vested interests in the outcome of the predictive model. Although performativity is generally problematic because it manifests as distribution shifts, we develop algorithmic fairness practices that leverage performativity to achieve stronger group fairness guarantees in social classification problems (compared to what is achievable in non-performative settings). In particular, we leverage the policymaker's ability to steer the population to remedy inequities in the long term. A crucial benefit of this approach is that it is possible to resolve the incompatibilities between conflicting group fairness definitions.

1:40 PM–2:00 PM Speaker: Jian Yan, Cornell University

Transfer learning for model-free variable importance

Author(s): Jian Yan, Cornell University; Charles Wolock, University of Pennsylvania; Yang Ning, Cornell University; Yong Chen, University of Pennsylvania

Quantifying variable importance in regression has received renewed interest lately. However, due to label scarcity and distribution shift occurring frequently in real-world applications, the data we collect may not be uniformly drawn from the population we are interested in. In this work, we aim to perform inference on a popular model-free variable importance measure in an unlabeled target population whose distribution differs from that of the labeled source data. We

propose a singly robust estimator, which is consistent regardless of whether the density ratio model is correctly specified or not. When the density ratio model is correctly specified, we establish the asymptotic normality of the proposed estimator and further show that it attains the semiparametric efficiency bound.

2:00 PM–2:20 PM Speaker: Yudong Wang, University of Pennsylvania

Integration of heterogeneous data via a one-shot distributed EM algorithm

Author(s): Yudong Wang, University of Pennsylvania; Xiaokang Liu, University of Missouri; Yang Ning, Cornell University; Raymond Carroll, Texas A&M University; Yong Chen, University of Pennsylvania

Integration of multi-site data while accounting for heterogeneity across sites is crucial in biomedical studies. To incorporate between-site heterogeneity in clustering analysis with multi-site data, we employ a heterogeneous mixture model where different sites share the same mixture components but with site-specific mixing proportions to account for the between-site heterogeneity. Due to regulatory restrictions, sharing patient-level data is commonly not allowed, which necessities distributed inference techniques for model fitting. However, distributed inference of this heterogeneous mixture model is challenging, and direct application of existing distributed algorithms may incur bias or prohibitive communication costs. In this study, we develop a novel one-shot distributed EM algorithm for fitting the heterogeneous mixture model. We show that our one-shot distributed estimator achieves the full-sample efficiency with only one round of communication of summary-level statistics across sites. Monte Carlo simulations and a real-data example in multisystem inflammatory syndrome are used to demonstrate the effectiveness of our method.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S47: ADVANCED STATISTICAL METHODS TO INFORM CLINICAL DECISION MAKING

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Southern Ground A/B (Mezzanine Level)

Organizer: Dandan Liu, Vanderbilt University Medical Center
Chair: Tianyi Sun, Vanderbilt University

1:00 PM–1:20 PM Speaker: Yuan Zhang, University of Pennsylvania

Dynamic risk assessment by landmark modeling of the restricted mean survival time

Author(s): Yuan Zhang, University of Pennsylvania; Douglas Schaubel, University of Pennsylvania

Dynamic assessment of the risk of adverse events is essential to informing treatment decisions, with the goal of optimizing patient outcomes and/or avoiding over-treatment. In liver transplantation, to ensure equitable allocation of available deceased-donor organs for end-stage liver disease patients, the risk of death is regularly assessed based on the change in biomarkers or vital signs to evaluate a patient's medical status and urgency of treatment receipt. In order to broaden the scope of jointly analyzing the longitudinal covariate process and the time-to-event outcome, we propose a landmark method which directly models the restricted mean survival time (RMST) while accommodating dependent censoring. Advantages of RMST models include removing the assumption that the hazard model (e.g., in a Cox regression) is correct at every time point, since RMST models consider a single restriction time as opposed to a process. Moreover, many investigators prefer the area under the survival curve over hazard rate as a clinical endpoint due to interpretability. The proposed methods are illustrated using national registry data on a cohort of patients wait-listed for a liver transplant.

1:20 PM–1:20 PM Speaker: Dandan Liu, Vanderbilt University Medical Center

Statistical challenges of implementing an EHR-based prediction model in real time

Author(s): Dandan Liu, Vanderbilt University Medical Center; Tianyi Sun, Vanderbilt University

Clinical prediction models have been widely acknowledged as informative tools providing evidence-based support for clinical decision making. However, such prediction models are often underused in clinical practice due to many reasons including the challenge of handling missing information upon real-time risk calculation and challenge of heterogeneity in standard of care across health care systems that impede adaption of existing prediction models. In this talk, I will showcase a real-world example of implementing a prediction model to identify emergency department patients with acute heart failure who can be safely discharged home rather than admitted to an inpatient setting, to demonstrate these challenges and associated opportunities for innovative statistical methods development and implementation.

1:40 PM–2:00 PM Speaker: Yuhao Deng, University of Michigan

Computationally efficient methods for estimating co-heritability of multivariate phenotypes using biobank data

Author(s): Yuhao Deng, University of Michigan; Yuanjia Wang, Columbia University; Donglin Zeng, University of Michigan

Biobank data provide a rich source to understand the degree of co-heritability among multiple disease phenotypes from shared genetic etiology. However, due to a large number of disease phenotypes of heterogenous types, estimating the co-heritability is both statistically and computationally challenging. In this work, we propose a joint model with latent polygenic effects for all the phenotypes, in which the random effects due to genetic etiology are modelled separately from the latent environmental effects. Computationally, we propose a two-stage procedure to first estimate the heritability and environmental correlation for a single phenotype, and then estimate the co-heritability between any two phenotypes by maximizing a pairwise pseudo-likelihood function. We extract nucleic family and apply divide-and-conquer approaches so that our algorithm can easily scale to analyzing the biobank data. Our numerical algorithms involve at most five-dimensional integration regardless of the number of the disease phenotypes, so are computationally efficient and reliable. Finally, the proposed method is illustrated through simulations and application to the UK biobank data.

2:00 PM–2:20 PM Speaker: Sergio Branciamore, City of Hope

Deciphering JAK/STAT pathway changes in breast cancer using Bayesian networks analysis

Author(s): Sergio Branciamore, City of Hope

Immune signaling networks (ISNs) play a pivotal role in health and disease, including breast cancer (BC). These intricate networks involve non-linear relationships among multiple variables. This complexity makes dissecting and modeling ISNs challenging, especially when using traditional statistical methods. In our study, we employed a Bayesian network reconstruction approach to explore how the ISN rewires itself in BC, with a specific focus on the JAK/STAT pathway. Our analysis extended beyond individual variable connections, allowing us to delve into the network's comprehensive interactions. This integrated approach provides a deeper understanding of the biological mechanisms at play. Our key findings include the description and characterization of distinct functional behaviors in different immunocell types. In particular, we studied non-classical monocytes and characterized various "programs." These "programs" are non-linear functions of PDL1 and FoxP3 that regulate the phosphorylation of pSTAT3 and pSTAT6. We related shifts in these programs to specific changes observed in BC patients. Finally, we used the belief propagation algorithm as a computational intervention method to study the simultaneous responses of multiple proteins and predict their responses to treatments such as interleukin administration. By addressing the complexity of the immune system, our computational approach could significantly enhance personalized treatment strategies for breast cancer patients.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S48: INCORPORATING MACHINE LEARNING COMPONENT IN STATISTICAL MODELING

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Blackbird B (Mezzanine Level)

Organizer: Song Zhang, University of Texas Southwestern Medical Center
Chair: Jing Cao, Southern Methodist University

1:00 PM–1:20 PM Speaker: Xinlei (Sherry) Wang, University of Texas at Arlington

Variational Bayesian semi-supervised keyword extraction

Author(s): Yaofang Hu, Southern Methodist University; Yichen Cheng, Georgia State University; Yuesen Xia, Georgia State University; Xinlei Wang, University of Texas at Arlington

The exponential expansion of textual data, stemming from various sources such as online product reviews and scholarly publications on scientific discoveries, has created an unprecedented demand for the extraction of succinct yet comprehensive information. As a result, in recent years, tremendous efforts have been spent in developing novel methodologies for keyword extraction. Although many methods have been proposed to automatically extract keywords in the contexts of both unsupervised and fully supervised learning, how to effectively utilize partially observed keywords, such as author-specified keywords and Twitter hashtags, remains an underexplored area. We propose a novel variational Bayesian semi-supervised (VBSS) keyword extraction approach, built on a recent Bayesian semi-supervised (BSS) technique that utilizes the information from a small set of known keywords in a document to identify previously undetected ones. Our proposed VBSS method greatly enhances the computational efficiency of BSS via mean-field variational inference, coupled with data augmentation, which brings closed-form solutions at each step of the optimization process. Further, our numerical results show that VBSS offers enhanced accuracy for long texts and improved control over false discovery rates when compared with a list of state-of-the-art keyword extraction methods.

1:20 PM–1:20 PM Speaker: Jing Cao, Southern Methodist University

Interpretable sentiment analysis using the attention-based multiple instance classification model

Author(s): Jing Cao, Chenyu Yang

Sentiment analysis (SA) is widely used for analyzing text data to identify the underlying opinion or emotion expressed in a document. Neural network methods in natural language processing have produced state-of-the-art results in SA. However, many of those methods are "black-box" algorithms which don't provide interpretable results. In this paper, we aim to develop a word-level context-based method with accessibility and interpretability. Specifically, we propose an attention-based multiple instance classification model (AMIC). AMIC uses word embedding to transform text to data, employs a multiple instance classification structure, and incorporates the self-attention mechanism to include context information from document. The accessibility comes from the fact that AMIC has a transparent model structure and it is easy to implement. The interpretability stems from AMIC's capability of providing word-level context-based sentiment metrics. In addition, AMIC can be used to construct domain-specific sentiment dictionary without requiring prior information on seed words or a pre-trained list of sentiment words. We demonstrate the performance of AMIC using a large online wine review dataset.

1:40 PM–2:00 PM Speaker: Yu Ryan Yue, Baruch College, The City University of New York

Scalable parameter-free Bayesian trend filtering on graphs

Author(s): Kamiar Rad, Baruch College, The City University of New York; Yu Ryan Yue, Baruch College, The City University of New York; Amanda Mejia, Indiana University

Latent trend estimation on graphs has been an active research area since the pioneering work in Leser (1961). Despite the vast related literature, various computational and statistical challenges remain unresolved. These challenges include scalability, spatial adaptivity, measures of uncertainty, and hyper-parameter tuning. In this paper, we address these challenges by proposing a fully Bayesian sampling methodology. Specifically, we fuse ideas such as trend filtering on graphs (TFG), global-local shrinkage priors (e.g., Laplace prior and horseshoe prior), and computationally efficient block-Gibbs sampling, to innovate a class of scalable Bayesian TFG models that lead to spatially adaptive, and parameter-free latent trends on graphs equipped with error-bars. Our extensive numerical experiments show the adaptivity and scalability of our approach. We further illustrate the usefulness of our proposed sampling method by an example of electrical recordings of spatially sensitive neurons (grid cells), and by an example of cortical surface functional magnetic resonance imaging.

2:00 PM–2:20 PM Speaker: Xiaowei Zhan, University of Texas Southwestern Medical Center

GNN-EGG: Graph Neural Network Explanations via Graph Generation

Author(s): Art Taychameekiatchai, Southern Methodist University; Xiaowei Zhan, University of Texas Southwestern Medical Center; Guanghua Xiao, University of Texas Southwestern Medical Center

Graph Neural Networks (GNNs) provide a means for modeling inherently graphical data, such as transportation, social, and molecular networks, but also for enhancing and reducing unstructured data, including text and images. A potential shortcoming is that their predictions are opaque—a "black box" hindering broader adoption and refinement. In this paper, we propose a novel architecture agnostic algorithm for generating model-level post-hoc explanations of GNN based graph classifiers. The algorithm attempts to learn the data generating distribution for each class of graph the model can predict. While this idea is not new, we introduce a novel loss function aimed at reducing the random baseline issue where completely random graphs can still yield similar embeddings and strong predictions. The primary contribution of this work is the use of a differentiable approximation to Graph Edit Distance in the loss function. This term enables us to ensure consistency in both the graph space and the embedding space for our representative examples. We illustrate the theoretical advantages of our loss function through a simulation study, benchmark our algorithm using the MUTAG dataset, and apply our method to a more complex GNN that predicts oral malignancies from pathology slides.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S49: RECENT DEVELOPMENTS FOR DYNAMIC AND TIME-TO-EVENT DATA ANALYSIS

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Gold (Lower Level)

Organizer: Dayu Sun, Indiana University
Chair: Dayu Sun, Indiana University

1:00 PM–1:20 PM Speaker: Dayu Sun, Indiana University

Semiparametric rate model for panel count data: Challenges of time-varying

Author(s): Dayu Sun, Indiana University

Panel count data are commonly encountered in event history studies and clinical trials, where covariates and coefficients often exhibit time-varying dynamics. Despite the prevalence of such data, existing statistical methodologies primarily rely on semiparametric mean models, which, constrained by monotonicity, encounter limitations. An alternative approach, the semiparametric rate model, better suited for time-varying covariates, remains underexplored due to theoretical and practical hurdles. Notably, challenges in deriving the explicit form of the least-favorable direction within semiparametric theory hinder robust variance estimation for hypothesis testing. To address these gaps, we introduce an EM-algorithm-based framework tailored for panel count data with time-varying covariates and coefficients. Several novel variance estimation methods are proposed. We establish a robust theoretical foundation and demonstrate its efficacy through rigorous numerical studies encompassing simulations and real-world applications, affirming its practical significance.

1:20 PM–1:20 PM Speaker: Ying Cui, Stanford University

Adaptive prediction strategy with individualized variable selection

Author(s): Bryan Cai, Stanford University; Ying Cui, Stanford University; Haoda Fu, Eli Lilly and Company; Donald M Lloyd-Jones, Northwestern University; Lihui Zhao, Northwestern University; Lu Tian, Stanford University

Today, physicians have access to a wide array of tests for diagnosing and prognosticating medical conditions. Ideally, they would apply a high-quality prediction model, utilizing all relevant features as input, to facilitate appropriate decision-making regarding treatment selection or risk assessment. However, not all features used in these prediction models are readily available to patients and physicians without incurring some costs. In practice, predictors are typically gathered as needed in a sequential manner, while the physician continually evaluates information dynamically. This process continues until sufficient information is acquired, and the physician gains reasonable confidence in making a decision. Importantly, the prospective information to collect may differ for each patient and depend on the predictor values already known. In this paper, we present a novel dynamic prediction rule designed to determine the optimal order of acquiring prediction features in predicting a clinical outcome of interest. The objective is to maximize prediction accuracy while minimizing the cost associated with measuring prediction features for individual subjects. To achieve this, we employ reinforcement learning, where the agent must decide on the best action at each step: either making a clinical decision with available information or continuing to collect new predictors based on the current state

of knowledge. To evaluate the efficacy of the proposed dynamic prediction strategy, extensive simulation studies have been conducted. Additionally, we provide two real data examples to illustrate the practical application of our method.

1:40 PM–2:00 PM Speaker: Xin He, University of Maryland

Semiparametric analysis of multivariate panel count data with informative observation processes
Author(s): Chang Chen, University of Maryland, College Park; Xin He, University of Maryland, College Park

Multivariate panel count data arise in studies involving several related types of recurrent events in which the study subjects are examined periodically over time. The observation times may vary from subject to subject and carry information about the underlying recurrent event processes of interest. In this paper, we propose a joint modeling approach to account for the informative observation processes using bivariate shared frailty models. Estimating equations and an EM algorithm are developed for the parameter estimation, and the resulting estimators are shown to be consistent and asymptotically normal. The proposed methods are evaluated through simulation studies and illustrated with an application to data from a skin cancer clinical trial.

2:00 PM–2:20 PM Speaker: Teng Fei, Memorial Sloan Kettering Cancer Center

Enhanced feature selection for microbiome data using FLORAL: Scalable log-ratio lasso regression
Author(s): Teng Fei, Memorial Sloan Kettering Cancer Center

Identifying predictive biomarkers of patient outcomes from high-throughput microbiome data is of high interest, while existing computational methods do not satisfactorily account for complex survival endpoints, longitudinal samples, and taxa-specific sequencing biases. We present FLORAL (https://vdblab.github.io/FLORAL/), an open-source computational tool to perform scalable log-ratio lasso regression and microbial feature selection for continuous, binary, time-to-event, and competing risk outcomes, with compatibility of longitudinal microbiome data as time-dependent covariates. The proposed method adapts the augmented Lagrangian algorithm for a zero-sum constraint optimization problem while enabling a two-stage screening process for extended false-positive control. In extensive simulation and real-data analyses, FLORAL achieved consistently better false-positive control compared to other lasso-based approaches, and better sensitivity over popular differential abundance testing methods for datasets with smaller sample size. In a survival analysis in allogeneic hematopoietic-cell transplant, we further demonstrated considerable improvement by FLORAL in microbial feature selection by utilizing longitudinal microbiome data over only using baseline microbiome data.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S50: ADDRESSING DESIGN AND ANALYSIS ISSUES IN PUBLIC HEALTH RESEARCH WITH COMPLICATED DATA STRUCTURES

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Melody (Lobby Level)

Organizer: Song Zhang, University of Texas Southwestern Medical Center
Chair: Dateng Li, Regeneron Pharmaceuticals

1:00 PM–1:20 PM Speaker: MinJae Lee, University of Texas Southwestern Medical Center

Statistical approaches to characterizing highly correlated measurements of health-related behaviors

Author(s): MinJae Lee, University of Texas Southwestern Medical Center

Promoting positive lifestyle behaviors (e.g., healthy diet, physical activity) to prevent and reduce the risk of cancer or related chronic diseases is a key focus of cancer prevention research. Given the growing population diversity, however, due to unobserved/undefined individual heterogeneity in multiple highly correlated measurements of behaviors, statistical modeling to assess these complex data is challenging. Biomarkers that identify high-risk individuals may improve our understanding of heterogeneity in risk behavioral patterns, but there is a lack of validated approach that can properly link biomarkers to multiple behaviors by determining their dynamic relations with cancer risk. This is because it requires a validation process and advanced statistical methodology that can address various challenges in analyzing biomarker data, including left-censoring due to detection limits. In this talk, I will introduce statistical approaches that can address these challenges and their applications in cancer prevention studies.

1:20 PM–1:20 PM Speaker: Dateng Li, Regeneron Pharmaceuticals

Sample size calculation for cluster randomized trials with zero-inflated count outcomes

Author(s): Zhengyang Zhou, University of North Texas Health Science Center; Dateng Li, Regeneron Pharmaceuticals; Song Zhang, University of Texas Southwestern Medical Center

Cluster randomized trials (CRT) have been widely employed in medical and public health research. Many clinical count outcomes, such as the number of falls in nursing homes, exhibit excessive zero values. In the presence of zero inflation, traditional power analysis methods for count data based on Poisson or negative binomial distribution may be inadequate. In this study, we present a sample size method for CRTs with zero-inflated count outcomes. It is developed based on GEE regression directly modeling the marginal mean of a zero-inflated Poisson outcome, which avoids the challenge of testing two intervention effects under traditional modeling approaches. A closed-form sample size formula is derived which properly accounts for zero inflation, ICCs due to clustering, unbalanced randomization, and variability in cluster size. Robust approaches, including t-distribution-based approximation and Jackknife re-sampling variance estimator, are employed to enhance trial properties under small sample sizes. Extensive simulations are conducted to evaluate the performance of the proposed method. An application example is presented in a real clinical trial setting.

1:40 PM–2:00 PM Speaker: Zhengyang Zhou, University of North Texas Health Science Center

Generalized estimating equations for analyzing alcohol outcomes for meta-analysis: A unified method
Author(s): Zhengyang Zhou, University of North Texas Health Science Center; Dateng Li; David Huh, University of Washington; Minge Xie, Rutgers University; Yong Chen, University of Pennsylvania; Justin Luningham, University of North Texas Health Science Center; E.Y. Mun, University of North Texas Health Science Center

Purpose: In alcohol research, many clinical endpoints are count variables, such as number of drinks consumed and number of alcohol-related negative consequences. Count outcomes may have distributional characteristics that impact the choice of statistical model, such as excessive zeros and overdispersion. Furthermore, these characteristics can vary depending on study-specific conditions. For example, universal interventions aimed at reducing alcohol consumption among all first-year college students may result in a disproportionate number of zero drinks observed due to inclusion of abstainers or infrequent drinkers, whereas indicated interventions for those at risk for heavy drinking may result in a small proportion of zeros. Such heterogeneity of data distributions across studies underscores a need for a flexible approach to analyzing count outcomes for research synthesis.

Methods: We developed a unifying, semiparametric statistical approach for count data that can handle various count outcome conditions in the context of meta-analysis, including (1) the proportion of zeros ranging from very few zeros in some studies to excessive zeros in others and (2) overdispersion. This unifying GEE approach for count outcomes is based on the generalized estimating equations (GEE) approach that directly models the overall mean of count outcomes and their variance. We compared the performance of the proposed GEE approach to other count outcome models in a simulation study across different intervention effects (small to medium effects) on alcohol consumption and different proportions of zero drinks (15% to 60%) in the data.

Results: In simulation studies, the proposed GEE approach had higher statistical power across conditions from very few zeros to excessive zeros, compared to competing methods, whereas the performance of other methods was compromised in certain conditions (e.g., power was 10% lower). In addition, the GEE approach had close to 100% convergence when estimating compared with 60% for other methods in some data conditions.

Conclusion: The proposed GEE approach provides a flexible, unifying analytical model for analyzing count outcomes that are commonly encountered in alcohol research. The approach can be particularly attractive for meta-analysis of individual participant data with count outcomes because this estimation provides a coherent analytical frame for combining data from heterogenous trials, including universal and indicated alcohol interventions. Support: R01 AA019511, K02 AA028630.

2:00 PM–2:20 PM Speaker: Song Zhang, University of Texas Southwestern Medical Center

A Bayesian adaptive design approach for stepped-wedge cluster randomized trials
Author(s): Jijia Wang, University of Texas (UT) Southwestern Medical Center; Jing Cao, Southern Methodist University; Chul Ahn, UT Southwestern Medical Center; Song Zhang, UT Southwestern Medical Center

The Bayesian group sequential design has been applied widely in clinical studies, especially in

Phase II and III studies. It allows early termination based on accumulating interim data. However, to date, there lacks development in its application to stepped-wedge cluster randomized trials, which are gaining popularity in pragmatic trials conducted by clinical and health care delivery researchers. We propose a Bayesian adaptive design approach for stepped-wedge cluster randomized trials, which makes adaptive decisions based on the predictive probability of declaring the intervention effective at the end of study given interim data. The Bayesian models and the algorithms for posterior inference and trial conduct are presented. We present how to determine design parameters through extensive simulations to achieve desired operational characteristics. We further evaluate how various design factors, such as the number of steps, cluster size, random variability in cluster size, and correlation structures, impact trial properties, including power, type I error, and the probability of early stopping. A application example is presented.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S51: THE JIANN-PING HSU INVITED SESSION ON BIOSTATISTICAL AND REGULATORY SCIENCES

Tuesday, June 18, 2024
1:00 PM–2:30 PM, Green Room (Lobby Level)

Organizer: Lili Yu, Georgia Southern University
Chair: Tolulope Adebile, Georgia Southern University

1:00 PM–1:20 PM Speaker: Tolulope Adebile, Georgia Southern University

Sleep duration and all-cause mortality in cancer patients: An association effect modified by hypertension.

Author(s): Tolulope Adebile, Jiann-Ping Hsu College of Public Health, Georgia Southern University; Purbasha Biswas, Jiann-Ping Hsu College of Public Health, Georgia Southern University; Olamide Asifat, Jiann-Ping Hsu College of Public Health, Georgia Southern University; Xinyan Zhang, Kennesaw State University; Jian Zhang, Jiann-Ping Hsu College of Public Health, Georgia Southern University; Lili Yu, Jiann-Ping Hsu College of Public Health, Georgia Southern University

Background: The relationship between sleep duration and hypertension in mortality prediction among the general population is not fully understood, including those with cancer. This study aims to investigate how sleep duration and hypertension interact to affect mortality risk in cancer patients.

Methods: We utilized data from 2,131 adults in the 2004 National Health Interview Survey linked to the 2019 National Death Index. Sleep duration was categorized into$\leq 6, 7-8, and \geq$9 hours. Survival probabilities were calculated using Kaplan-Meier methods, while Cox proportional hazard models provided estimates for univariate and multivariate-adjusted hazard ratios (HRs) and 95% confidence intervals (CIs). The potential role of hypertension as an effect modifier of the association between mortality and sleep duration was examined.

Results: Hypertension does not act as an effect modifier in the mortality-sleep duration relationship (p > 0.05). Based on the Kaplan-Meier estimates, the lowest survival probabilities occurred at 9 hours for cancer patients, with hypertensive participants showing generally worse mortality outcomes. In the fully adjusted stratified model, significant HRs were found only among normotensive individuals at≥ 9 hours [HR: 1.413, 95% CI: 1.067, 1.871]), compared to 7-8 hours. Contrarily, lower but non-significant HRs were observed at short sleep durations (≤ 6 hours) for hypertensive and normotensive individuals, compared to 7-8 hours.

Conclusion: The findings underscore the importance of adopting personalized cancer management strategies that consider an individual's hypertension status and sleep habits. Hypertension does not significantly alter the relationship between sleep duration and mortality in individuals with cancer. Notwithstanding, extended sleep durations of ≥ 9 hours may elevate mortality risks among normotensive cancer patients. Further research is warranted to delve deeper into these complex interactions, especially among hypertensive individuals.

1:20 PM–1:20 PM Speaker: Ibrahim Alliu, Georgia Southern University

Optimizing sample size for accelerated failure time model in progressive type-ii censoring through ranked set sampling

Author(s): Ibrahim Alliu, Jiann-Ping Hsu College of Public Health, Georgia Southern University, USA; Lili Yu, Jiann-Ping Hsu College of Public Health, Georgia Southern University, USA; Jing Kersey, Jiann-Ping Hsu College of Public Health, Georgia Southern University, USA; Hani Samawi, Jiann-Ping Hsu College of Public Health, Georgia Southern University, USA

Survival data is a type of data that arises when the time from a defined time until the occurrence of a particular event, such as time to death from small cell lung cancer after diagnosis, length of time in remission for leukemia patients, and length of stay (i.e., time until discharge) in hospital after surgery. The accelerated failure time (AFT) models are popular linear models to analyze survival data. It provides a linear relationship between the log of the failure time and covariates that affect the expected failure time by contracting or expanding the time scale. This paper examines the performance of the Ranked Set Sampling (RSS) scheme on the AFT models for Progressive Type-II censoring-survival data. The RSS scheme is a sampling scheme that selects a sample based on a baseline auxiliary variable for assessing survival time. Simulation studies show that this approach provides a more robust testing procedure and a more efficient hazard ratio estimate than simple random sampling (SRS). The lung cancer survival data are used to demonstrate the method.

1:40 PM–2:00 PM Speaker: Varadan Sevilimedu, Memorial Sloan Kettering Cancer Center

An improved MC-SIMEX algorithm

Author(s): Varadan Sevilimedu, Memorial Sloan Kettering Cancer Center; Lili Yu, Jiann Ping Hsu College of Public Health, Georgia Southern University

Misclassification Simulation-Extrapolation (MC-SIMEX) is an established method to correct for misclassification in binary covariates in a model. It involves the use of simulation component which simulates pseudo-datasets with added degree of misclassification in the binary covariate and an extrapolation component which models the covariate's regression coefficients obtained at each level of misclassification using a quadratic function and extrapolates it to a point of "no error" in the classification of the binary covariate under question. However, extrapolation functions are not usually known accurately beforehand and are therefore only approximated versions. In this study, we propose an innovative method that eliminates the need for an extrapolation function through the use of a derived relationship between the naive regression coefficient estimates and the true regression coefficient. Simulation studies are conducted to study and compare the numerical properties of the resulting estimator to the original MC-SIMEX estimator. Real data analysis is also provided.

2:00 PM–2:20 PM Speaker: Chang Wang, University of Michigan

Jiann-Ping Hsu Pharmaceutical and Regulatory Sciences Student Paper Award Winner: Adaptive weight learning for multiple outcome optimization with continuous treatment

Author(s): Chang Wang, University of Michigan; Lu Wang, University of Michigan

To promote precision medicine, individualized treatment regimes (ITRs) are crucial for optimizing the expected clinical outcome based on patient-specific characteristics. However, existing

ITR research has primarily focused on scenarios with categorical treatment options and a single outcome. In reality, clinicians often encounter scenarios with continuous treatment options and multiple, potentially competing outcomes, such as medicine efficacy and unavoidable toxicity. To balance these outcomes, a proper weight is necessary, which should be learned in a data-driven manner that considers both patient preference and clinician expertise. In this paper, we present a novel algorithm for developing individualized treatment regimes (ITRs) that incorporate continuous treatment options and multiple outcomes, utilizing observational data. Our approach assumes that clinicians are optimizing individualized patient utilities with sub-optimal treatment decisions that are at least better than random assignment. Treatment assignment is assumed to directly depend on the true underlying utility of the treatment rather than patient characteristics. The proposed method simultaneously estimates the weighting of multiple outcomes and the decision-making process, allowing for construction of individualized treatment regimes with continuous doses. The proposed estimator can be used for inference and variable selection, facilitating the identification of informative treatment assignments and preference-associated variables. We evaluate the finite sample performance of our proposed method via simulation studies and apply it to a real data application of radiation oncology analysis.

2:20 PM–2:30 PM **Q&A and Floor Discussion**

S52: STATISTICS IN BIOSCIENCES (SIBS): RECENT METHODOLOGICAL ADVANCES FOR TRANSFORMING DATA INTO KNOWLEDGE

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Symphony 1 (Lobby Level)

Organizer: Hongkai Ji, Johns Hopkins University
Chair: Hongkai Ji, Johns Hopkins University

3:00 PM–3:20 PM Speaker: Jia Zhao, Yale University

scPI: A scalable framework for probabilistic inference in single-cell RNA-sequencing data analysis

Author(s): Jingsi Ming, East China Normal University; Jia Zhao, Yale University; Can Yang, Hong Kong University of Science and Technology

The technique of single-cell RNA-sequencing (scRNA-seq) has provided an unprecedented opportunity to investigate the cellular heterogeneity of complex tissues. As large-scale scRNA-seq datasets are becoming more available and affordable, there is a growing demand for computational scalable methods to analyze scRNA-seq data. Here, we propose a scalable framework, scPI, to infer the latent low-dimensional representations of the scRNA-seq data to facilitate downstream analysis. Our method scPI makes use of the amortized variational inference, where the posterior mean and variance of the latent variable are parameterized by a nonlinear neural network. This inference structure combined with stochastic optimization enables its computational efficiency and scalability. Through the analysis of two real datasets, we demonstrate that the scPI framework can be effectively applied to several probabilistic models for scRNA-seq data, in terms of its scalability, missing value imputation and cell type clustering.

3:20 PM–3:20 PM Speaker: Zhezhen Jin, Columbia University

A step-wise multiple testing for linear regression models with application to the study of resting energy expenditure

Author(s): Junyi Zhang, Baruch College; Zimian Wang, Columbia University; Zhezhen Jin, Columbia University; Zhiliang Ying, Columbia University

Motivated by the mechanistic model of the resting energy expenditure, we present a new multiple hypothesis testing approach to evaluate organ/tissue-specific resting metabolic rates. The approach is based on generalized marginal regression estimates for a subset of coefficients along with a stepwise multiple testing procedure with a minimization–maximization of the normalized estimates (maximization over all its components and minimization over all possible choices of the subset). The approach offers a valid way to address challenges in multiple hypothesis testing on regression coefficients in linear regression analysis especially when covariates are highly correlated. Importantly, the approach yields estimates that are conditionally unbiased. In addition, the approach controls a family-wise error rate in the strong sense. The approach was used to analyze a real study on resting energy expenditure in 131 healthy adults, which yielded an interesting and surprising result of age-related decrease in resting metabolic rate of kidneys. Simulation studies were also presented with various strengths of multi-collinearity induced by pre-specified correlation in covariates.

3:40 PM–4:00 PM Speaker: Qingning Zhou, University of North Carolina at Charlotte

Analysis of the Cox model with longitudinal covariates with measurement errors and partly interval censored failure times, with application to an AIDS clinical trial

Author(s): Yanqing Sun, University of North Carolina at Charlotte; Qingning Zhou, University of North Carolina at Charlotte; Peter Gilbert, Fred Hutchinson Cancer Center

Time-dependent covariates are often measured intermittently and with measurement errors. Motivated by the AIDS Clinical Trials Group (ACTG) 175 trial, this paper develops statistical inferences for the Cox model for partly interval censored failure times and longitudinal covariates with measurement errors. The conditional score methods developed for the Cox model with measurement errors and right censored data are no longer applicable to interval censored data. Assuming an additive measurement error model for a longitudinal covariate, we propose a nonparametric maximum likelihood estimation approach by deriving the measurement error induced hazard model that shows the attenuating effect of using the plug-in estimate for the true underlying longitudinal covariate. An EM algorithm is devised to facilitate maximum likelihood estimation that accounts for the partly interval censored failure times. The proposed methods can accommodate different numbers of replicates for different individuals and at different times. Simulation studies show that the proposed methods perform well with satisfactory finite-sample performances and that the naive methods ignoring measurement error or using the plug-in estimate can yield large biases. A hypothesis testing procedure for the measurement error model is proposed. The proposed methods are applied to the ACTG 175 trial to assess the associations of treatment arm and time-dependent CD4 cell count on the composite clinical endpoint of AIDS or death.

4:00 PM–4:30 PM **Q&A and Floor Discussion**

S53: DIAGNOSTIC ACCURACY AND BIOMARKER ANALYSIS

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Blackbird A (Mezzanine Level)

Organizer: Zhen Chen, National Institute of Child Health and Human Development (NICHD), NIH
Chair: Ruijin Lu, Washington University in St. Louis

3:00 PM–3:20 PM Speaker: Ying Huang, Fred Hutchinson Cancer Center

Assessing screening efficacy in the presence of cancer overdiagnosis

Author(s): Ying Huang, Fred Hutchinson Cancer Center; Ziding Feng, Fred Hutchinson Cancer Center

Cancer screening facilitates the early detection of cancer, at a stage when treatment is often most effective. However, it also brings the risk of overdiagnosis, where a diagnosis made through screening would not have led to symptoms or death during the patient's lifetime. In this project, we tackle a significant unresolved issue in the evaluation of screening efficacy: selecting primary endpoints and inferential procedures that efficiently consider potential overdiagnosis in screening trials. This is motivated by the necessity to design and analyze a phase IV Early Detection Initiative (EDI) trial for evaluating a pancreatic cancer screening strategy. We introduce two novel approaches for assessing screening efficacy, grounded on cancer stage-shift. These methods address potential overdiagnosis by: i) borrowing information about clinical diagnosis from the control arm that hasn't undergone screening (the BR approach), and ii) performing sensitivity analysis, contingent upon a conservative bound of the overdiagnosis magnitude (the SEN-T approach). Analytical methods and extensive simulation studies underscore the superiority of our proposed methods, demonstrating enhanced efficiency in estimating and testing screening efficacy compared to existing methods. The latter either overlook overdiagnosis or adhere to a valid, yet conservative, cumulative incidence endpoint. We illustrate the practical application of these approaches using ovarian cancer data from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. The results affirm that our methods bolster an efficient and robust study design for cancer screening trials.

3:20 PM–3:20 PM Speaker: Danping Liu, National Cancer Institute, NIH

Evaluation of test reproducibility using model without gold standard

Author(s): Danping Liu, National Cancer Institute, NIH

Monitoring the reproducibility of a diagnostic test is a critical aspect of quality control to ensure the reliability and validity of scientific results. A common practice is to use test-retest data to estimate agreement measures. However, when the disease prevalence is low, these measures can become overly sensitive to prevalence levels, making it challenging to establish a reasonable threshold. In this context, we propose a novel procedure of evaluating the test reproducibility, based on the estimation of test sensitivity and specificity over a range of prevalence levels. The criterion for determining insufficient reproducibility is the identification of a prevalence level where both estimated sensitivity and specificity are simultaneously lower than the desired levels. With extensive simulation studies, we demonstrate that the proposed test procedure maintains a nominal type 1 error and attains high statistical power to detect reductions in accuracy, and hence can serve as a good quality control tool for assay monitoring. Our work is motivated by and applied to the quality control experiment of TypeSeq2, a high-throughput and low-cost assay for detecting HPV genotypes.

3:40 PM–4:00 PM Speaker: Jin Yang, National Institute of Child Health and Human Development

Youden index estimation based on group-tested data

Author(s): Jin Yang, National Institute of Child Health and Human Development (NICHD), NIH; Aiyi Liu, NICHD; Neil J. Perkins, NICHD; Zhen Chen, NICHD

Youden index, a linear function of sensitivity and specificity, provides a direct measurement of the highest diagnostic accuracy achievable by a biomarker. It is maximized at the cut-off point that optimizes the biomarker's overall classification rate while assigning equal weight to sensitivity and specificity. In this paper, we consider the problem of estimating the Youden index when only group-tested data are available. The unavailability of individual disease statuses poses a challenge, especially when there is differential false positives and negatives in disease screening. We propose both parametric and nonparametric procedures for estimation of the Youden index and exemplify our methods by utilizing data from the National Health and Nutrition Examination Survey (NHANES) to evaluate the diagnostic ability of monocyte for predicting chlamydia.

4:00 PM–4:20 PM Speaker: Soutik Ghosal, University of Virginia

A placement value-based ROC surface spline regression

Author(s): Soutik Ghosal, University of Virginia

Receiver Operating Characteristic (ROC) curves are a popular tool for assessing the diagnostic accuracy of biomarkers when predicting binary disease outcomes. However, this method becomes limited when dealing with more than two outcome groups; it cannot simultaneously distinguish among them. In these cases, ROC surfaces (for three outcome groups) or ROC manifolds (for more than three outcome groups) are more suitable approaches. Additionally, biomarker performance can be influenced by various covariates. Incorporating these covariates into the analysis can enhance the diagnostic accuracy model and provide insights into the biomarker's performance across different covariate levels. Previous approaches to ROC surface regression have typically involved modeling biomarkers for each outcome group against their respective covariates or using ordinal/multinomial logistic regression. While useful, these methods might not capture the full complexity of covariate influences. In this project, we propose a novel approach to ROC surface regression using Placement Value (PV) to address the nonlinear impact of covariates. PV standardizes diseased biomarker groups relative to the healthy group, providing a more direct way to model the nonlinear effects of covariates. This approach establishes the ROC surface as a function of two sets of placement values, enabling us to account for nonlinear covariate impacts directly. Moreover, this method allows us to derive covariate-specific optimal thresholds, enabling better discrimination of outcomes based on those covariate levels. We plan to apply our method to the National Institute of Child Health and Human Development (NICHD) Fetal Growth Studies data to evaluate the diagnostic performance of biomarkers in predicting abnormal birthweight outcomes, such as small- and large-for-gestational-age infants, at different gestation periods.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S54: EMPOWERING CLINICAL TRIAL EFFICIENCY THROUGH UTILIZATION OF REAL-WORLD DATA: INNOVATIONS AND APPLICATIONS

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Lyric (Lobby Level)

Organizer: Rachael Liu, Takeda Pharmaceuticals
Chair: Rachael Liu, Takeda Pharmaceuticals

3:00 PM–3:20 PM Speaker: Jianchang Lin, Takeda Pharmaceuticals

Incorporating external real-world data for hybrid confirmatory adaptive design

Author(s): Jianchang Lin, Takeda Pharmaceuticals; Rachael Liu, Takeda Pharmaceuticals; Junjing Lin, Takeda Pharmaceuticals

Adaptive designs, such as group sequential designs (and the ones with additional adaptive features) or adaptive platform trials, have been quintessential efficient design strategies in trials of unmet medical needs, especially for generating evidence from global regions. Such designs allow interim decision making and making adjustment to study design when necessary, meanwhile maintaining study integrity and operating characteristics. However, driven by the heightened competitive landscape and the desire to bring effective treatment to patients faster, innovation in the already functional designs is still germane to further propel drug development to a more efficient path. One way to achieve this is by leveraging external real-world data (RWD) in the adaptive designs to support interim or final decision making. In this paper, we propose a novel framework of incorporating external RWD in adaptive design to improve interim and/or final analysis decision making. Within this framework, researchers can prespecify the decision process and choose the timing and amount of borrowing while maintaining objectivity and controlling of type I error. Simulation studies in various scenarios are provided to describe power, type I error, and other performance metrics for interim/final decision making. A case study in non-small cell lung cancer is used for illustration on proposed design framework.

3:20 PM–3:20 PM Speaker: Ying Yuan, University of Texas MD Anderson Cancer Center

Self-adapting mixture prior to dynamically borrow information from historical data in clinical trials

Author(s): Ying Yuan, University of Texas MD Anderson Cancer Center

Mixture priors provide an intuitive way to incorporate historical data while accounting for potential prior-data conflict by combining an informative prior with a non-informative prior. However, pre-specifying the mixing weight for each component remains a crucial challenge. Ideally, the mixing weight should reflect the degree of prior-data conflict, which is often unknown beforehand, posing a significant obstacle to the application and acceptance of mixture priors. To address this challenge, we introduce self-adapting mixture (SAM) priors that determine the mixing weight using likelihood ratio test statistics. SAM priors are data-driven and self-adapting, favoring the informative (non-informative) prior component when there is little (substantial) evidence of prior-data conflict. Consequently, SAM priors achieve dynamic information borrowing. We demonstrate that SAM priors exhibit desirable properties in both finite and large samples and achieve information-borrowing consistency. Moreover, SAM priors are easy to compute, data-driven, and calibration-free, mitigating the risk of data dredging. Numer-

ical studies show that SAM priors outperform existing methods in adopting prior-data conflicts effectively. We developed R package SAMprior and web application that are freely available at CRAN and www.trialdesign.org to facilitate the use of SAM priors.

3:40 PM–4:00 PM Speaker: Kentaro Takeda, Astellas

Hybrid control design with commensurate prior constructed using propensity score-matched external controls

Author(s): Kentaro Takeda, Astellas; Yusuke Yamaguchi, Astellas

Randomized controlled trials (RCTs) are a gold standard for demonstrating substantial evidence of the effectiveness of a new drug for regulatory approval, though there are some special situations where the conduct of RCTs is unfeasible due to the small patient population. A hybrid control design is a way to borrow information from external controls to augment concurrent controls in the RCT and is expected to overcome the feasibility issue when adequate RCTs cannot be conducted. A major challenge in the hybrid control design is its inability to eliminate a prior-data conflict caused by systematic imbalances in measured or unmeasured confounders between patients in the concurrent treatment/control group and external controls. To prevent the prior-data conflict, combined use of propensity score matching and Bayesian commensurate prior has been proposed in a hybrid control design context. The propensity score matching is first performed to guarantee the balance in baseline characteristics, and then the Bayesian commensurate prior is constructed while discounting the information based on the similarity in outcomes between the concurrent and external controls. The process is straightforward, but some implementation aspects should be fully addressed. For example, it needs to be clarified which concurrent groups to use for matching (concurrent treatment, concurrent control, or both). In addition, the necessity of unequal matching would add further complexity. We conducted a comprehensive simulation study to evaluate the operating characteristics of the design, revealing some key features to be careful in the implementation.

4:00 PM–4:20 PM Speaker: Thomas Jemielita, Merck & Co.

Augmented RCTs: Mitigating bias when using external information

Author(s): Thomas Jemelita, Merck & Co.

Following the 21st-century Cures Act, there has been increasing interest in using real-world data (RWD) to enhance decision making for clinical development. For example, RWD could be used to benchmark a single-arm study or augment an existing control arm. While there are promising applications, there are considerable challenges such as data-source selection and confounding due to using nonrandomized data. One way to minimize the impact of potential confounding is the Hybrid RCT design, which randomizes R:1 between the study treatment and the control and then augments the RCT control with RWD control patients. Compared to a single arm design, the Hybrid RCT can use novel methods such as partial bias correction or simultaneously leverage methods such as propensity weighting and Bayesian information borrowing; either approach can better mitigate biased treatment estimates. Even with these benefits, RWD patients must also satisfy the same eligibility criteria as the RCT participants. This is challenging due to missing data, along with the presence of criteria such as "life expectancy" which is not directly observable but is used in most oncology clinical trials. Here, we propose a generalized propensity score weighting procedure to adjust for the life expectancy

requirement, which can further reduce potential bias due to unmeasured confounding. Simulation studies and real data examples with EHR-based and registry-based external control data will be used to highlight key concepts.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S55: RELIABLE AND RIGOROUS INFERENCE FOR BRAIN STRUCTURE AND NETWORKS

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Ocean Way (Mezzanine Level)

Organizer: Kaidi Kang, Vanderbilt University Medical Center
Chair: Simon Vandekar, Vanderbilt University Medical Center

3:00 PM–3:20 PM Speaker: Joshua Lukemire, Emory University

Reliable identification of covariate-related differences in brain functional networks

Author(s): Joshua Lukemire, Emory University; Ying Guo, Emory University

We introduce a general framework for decomposing brain function into functional brain networks for multi-subject data with repeated measurements and covariate effects. This general method provides a rigorous and much needed tool for investigating brain networks and their differences in imaging studies with complex study designs including longitudinal and/or multi-center studies. Our approach incorporates effects corresponding to data collection sites to correct for site-level biases and incorporates subject-specific effects to accommodate within-subject repeated measures such as those from longitudinal studies. Through simulations, we show that the proposed method has considerably improved performance as compared to other potential blind source separation approaches. We apply our procedure to study aging in the longitudinal ABCD study data and demonstrate functional brain network differences that supplement previous work showing age and sex related differences.

3:20 PM–3:20 PM Speaker: Mikail Rubinov, Vanderbilt University

The problem of overspecification in systems neuroscience

Author(s): Aditya Nanda, Vanderbilt University; Alireza Abbasi, Vanderbilt University; Mikail Rubinov, Vanderbilt University

Systems neuroscience is witnessing continuous improvements in the quality of data but lacks similar improvements in conceptual understanding. This mismatch arises, in part, from the difficulty of formalizing neuroscientific knowledge and rigorously testing new discovery against this knowledge. Here, we define this problem as a problem of overspecification and contrast it with the more commonly known problem of overfitting. We discuss that overspecification confounds key results and affects many studies in the field. Finally, we describe solutions for integrating knowledge of large-scale brain activity and connectivity to help resolve this problem in systems neuroscience.

3:40 PM–4:00 PM Speaker: Kaidi Kang, Vanderbilt University

Study design features that improve the effect sizes in brain-wide association studies

Author(s): Kaidi Kang, Vanderbilt University

Several recent studies have raised concerns about the replicability of brain-wide association studies (BWAS). Here, we perform analyses and meta-analyses of a robust effect size index using 63 longitudinal and cross-sectional MRI studies (77,695 total scans) to demonstrate that optimizing study design is an important way to improve standardized effect sizes and replica-

bility in BWAS. A meta-analysis of brain volume associations with age indicates that BWAS with larger covariate variance have larger effect size estimates and that the longitudinal studies we examined have systematically larger standardized effect sizes than cross-sectional studies. Analyzing age effects on global and regional brain measures in the Lifespan Brain Chart Consortium, we show that modifying longitudinal study design to increase between-subject variability and adding a single additional longitudinal measurement per subject improves effect sizes. However, evaluating these longitudinal sampling schemes on cognitive, psychopathology, and demographic associations with structural and functional brain outcome measures in the Adolescent Brain and Cognitive Development dataset shows that longitudinal studies can, counterintuitively, be detrimental to effect sizes. We demonstrate that the benefit of conducting longitudinal studies depends on the strengths of the between- and within-subject associations of the brain and non-brain measures. Explicitly modeling between- and within-subject effects avoids conflating the effects and allows optimizing effect sizes for them separately. These findings underscore the importance of considering design features in BWAS and emphasize that increasing sample size is not the only approach to improve the replicability of BWAS.

4:00 PM–4:20 PM Speaker: Bennett Landman, Vanderbilt University

Beautiful and strange: Data-driven inference with diffusion MRI
Author(s): Bennett Landman, Vanderbilt University

Diffusion-weighted MRI (dMRI) is a unique imaging technique used to study the microstructural properties of tissues, particularly in the brain. Traditionally dMRI relies on predefined models, which may not fully capture the complexity of the underlying biological processes. This talk explores data-driven methods that enhance the inference capabilities in dMRI, leveraging advances in artificial intelligence (AI) and deep learning. We present data drive approaches from both the microstructural (e.g., voxel-wise diffusivity properties) and macrostructural (e.g., tractography) perspectives. We find that data-driven methods, particularly those leveraging deep learning, offer significant advantages in analyzing diffusion MRI data. By integrating data-driven information, we can achieve more accurate and reliable estimators with specifically tailored properties.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S56: STATISTICAL ADVANCES FOR COMPLEX EVENT HISTORY DATA

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Platinum (Lower Level)

Organizer: Wenbo Wu, New York University Grossman School of Medicine
Chair: Richard Liu, New York University Grossman School of Medicine

3:00 PM–3:20 PM Speaker: Mengling Liu, New York University

Dynamic single-index Cox regression model

Author(s): Yiwei Li, New York University (NYU) Grossman School of Medicine; Yuyan Wang, NYU Grossman School of Medicine; Mengling Liu, NYU Grossman School of Medicine

In examining multiple time-dependent exposures with time-to-event outcomes, the classical Cox regression model is limited in use due to its strong assumptions of linearity and constant hazard ratios. To bridge this gap, we propose a novel Partial Linear Dynamic Single-Index Cox regression model, which combines the time-varying impact of exposure on survival risk through a nonparametric single-index function with the linear effects of additional covariates. We employed regression spline tensor bases to approximate the single-index function and proposed a profile optimization algorithm to estimate the model. We also present a nonparametric test to formally evaluate the exposure mixture's temporal effect on survival risk. After establishing the large sample properties for the proposed estimator, we evaluate its finite-sample performance under extensive simulation scenarios and real data applications.

3:20 PM–3:20 PM Speaker: Yongzhao Shao, New York University Grossman School of Medicine

New survival models of risk factors for age at diagnosis of Alzheimer's disease and age at death
Author(s): Yongzhao Shao, Qiao Zhang

Alzheimer's disease (AD) is currently a leading cause of death without a cure. In the study of the effect of risk factors on mortality related to late-onset Alzheimer's disease, ignoring the presence of major competing causes of death (e.g. death due to cardiovascular disease or cancer) can lead to severe biases. We conduct survival analyses of the potentially heterogeneous effects of risk factors on the age at diagnosis of AD and age at death accounting for the presence of competing causes of death. The proposed model contains the mixture cure models as special cases. As an application, we assess the heterogeneous effects of the allele 4 of the APOE gene on AD-related mortality in the context of competing risk using the National Alzheimer's Coordinating Center (NACC) database. This talk is based on joint research with Qiao Zhang and Elizabeth Pirraglia at New York University.

3:40 PM–4:00 PM Speaker: Yuxiang Wu, University of Missouri

Group variable selection for Cox model with interval-censored failure time data
Author(s): Yuxiang Wu, University of Missouri; Hui Zhao, Zhongnan University of Economics and Law; Jianguo Sun, University of Missouri

Group variable selection is often required in many areas and for this, many methods have been developed under various situations. Unlike the individual variable selection, the group variable selection can select the variables in groups and it is more efficient to identify both important

and unimportant variables or factors by taking into account the existing group structure. In this paper, we consider the situation where one only observes interval-censored failure time data arising from Cox model, for which there does not seem to exist an established method. More specifically, a penalized sieve maximum likelihood selection and estimation procedure is proposed and the oracle property of the proposed method is established. Also, an extensive simulation study is performed and suggests that the proposed approach works well in practical situations. An application of the method to a set of real data is provided.

4:00 PM–4:20 PM Speaker: Wenbo Wu, New York University Grossman School of Medicine

Competing risk modeling with bivariate varying coefficients to understand the dynamic impact of COVID-19

Author(s): Wenbo Wu, New York University Grossman School of Medicine; John Kalbfleisch, University of Michigan; Jeremy Taylor, University of Michigan; Jian Kang, University of Michigan; Kevin He, University of Michigan

The coronavirus disease 2019 (COVID-19) pandemic has exerted a profound impact on patients with end-stage renal disease relying on kidney dialysis to sustain their lives. A preliminary analysis of dialysis patient postdischarge hospital readmissions and deaths in 2020 revealed that the COVID-19 effect has varied significantly with postdischarge time and time since the pandemic onset. However, the complex dynamics cannot be characterized by existing varying coefficient models. To address this issue, we propose a bivariate varying coefficient model for competing risks, where tensor-product B-splines are used to estimate the surface of the COVID-19 effect. An efficient proximal Newton algorithm is developed to facilitate the fitting of the new model to the massive data for Medicare beneficiaries on dialysis. Difference-based anisotropic penalization is introduced to mitigate model overfitting and effect wiggliness; a cross-validation method is derived to determine optimal tuning parameters. Hypothesis testing procedures are designed to examine whether the COVID-19 effect varies significantly with postdischarge time and the time since the pandemic onset, either jointly or separately. Applications to Medicare dialysis patients demonstrate the real-world performance of the proposed methods. Simulation experiments are conducted to evaluate the estimation accuracy, type I error rate, statistical power, and model selection procedures. Supplementary materials for this article are available online.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S57: REAL-WORLD EVIDENCE IN DRUG DEVELOPMENT

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Sound Emporium A/B (Mezzanine Level)

Organizer: Gang Li, Eisai
Chair: Xiaolong Luo, Sarepta Therapeutics

3:00 PM–3:20 PM Speaker: Tong Li, Sanofi

Generalizing treatment effect to a target population without individual patient data

Author(s): Hui Quan, Sanofi; Tong Li, Sanofi; Xun Chen, Sanofi; Gang Li, Eisai

The innovative use of real-world data (RWD) can answer questions that cannot be addressed using data from randomized clinical trials (RCTs). However, many methods for analyzing data from RCTs cannot be directly applied and the usage of causal inference framework is required for RWD analysis. In addition, to protect patient privacy, sharing individual patient data (IPD) is subject to strict regulations and is logistically prohibitive. In this research, we propose using a double inverse probability weighting (DIPW) approach to estimate the population average treatment effect (PATE) for a target population without the need for the analysis sponsor to access IPD. One probability weighting is for patient matching to ensure the patient comparability across treatments. Another probability weighting is for generalizing the result from a subpopulation of patient with data of the endpoint to the whole target population. The likelihood expressions for propensity scores and the DIPW estimator of the PATE can be rewritten to only rely on regional summary statistics that do not require access to IPD. Our approach hinges upon the positivity and conditional independency assumptions, prerequisites to most RWD analysis approaches. Simulations are conducted to compare the performances of the proposed method against a modified meta-analysis and a regular meta-analysis.

3:20 PM–3:20 PM Speaker: Zailong Wang, Vertex Pharmaceuticals

Matching adjusted indirect comparison: Methods and applications review

Author(s): Zailong Wang, Vertex Pharmaceuticals

In a setting that individual patient data for the treatment group are available from clinical trials without direct comparator and the comparator's aggregated data are available from the published manuscripts, to adjust for between-trial differences in the distribution of the relevant baseline variables that influence the outcome, matching adjusted indirect comparison (MAIC) is recommended for the indirect treatment comparisons. This presentation will illustrate MAIC methods and applications. The process for MAIC includes comparator data choices, matching variables selection, weight calculation, and outcome comparisons for different scenario of endpoints. Example applications from publications to illustrate how indirect comparisons based only on aggregate data can be limited by cross-trial differences in patient populations, differences in the definitions of outcome measures, and sensitivity to modeling assumptions will also be presented.

3:40 PM–4:00 PM Speaker: Xiaolong Luo, Sarepta Therapeutics

Topics in application of RWD for rare disease drug development

Author(s): Xiaolong Luo, Sarepta Therapeutics; Xiao Ni, Sarepta Therapeutics; Junming Zhu, Sarepta Therapeutics; Jing Li, Sarepta Therapeutics; Kai Ding, Sarepta Therapeutics; Weihua Tang, Sarepta Therapeutics

Real-world data (RWD) has become indispensable in drug development, particularly for patients with rare diseases where traditional patient data and precedents are scarce. It serves pivotal roles in modeling disease progression, defining patient population, and assessing treatment outcomes. However, challenges persist in ensuring the quality, privacy, transparency, and standardization of RWD. These challenges often manifest as missing data issues, which, if not meticulously addressed, can lead to substantial bias. For example, in a degenerative disease, endpoints sometimes were no longer collected after disease progression, hence the observed longer-term data may reflect slower progression from an enriched healthier group. External control comparison using observed RWD data could only yield an overly conservative treatment effect estimate. While regulatory bodies and Good Clinical Practice (GCP) stakeholders such as the International Council for Harmonisation (ICH), the Food and Drug Administration (FDA), the National Institutes of Health (NIH), and various private and peer-reviewed publications have provided guidelines for managing missing data, the effectiveness of these approaches varies case by case.

In this presentation, we will explore illustrative examples and demonstrate how scientifically and clinically sensible data handing techniques, including systematic data harmonization and imputation, can substantially reduce bias in RWD due to missing data. These methods provide rare disease drug developers the ability of performing more comprehensive analyses using robustly processed RWD, which can effectively uncover treatment effects that might otherwise remain obscured within the noise of imperfectly collected RWD.

4:00 PM–4:20 PM Speaker: Sai Dharmarajan, Sarepta Therapeutics

Quantifying survival benefit in real-world settings: Choice of time scale, time zero and effect measure

Author(s): Sai Dharmarajan, Sarepta Therapeutics

For survival or time-to-event endpoints, oftentimes the quantification of the treatment effect can proceed in many ways. In real-world settings, since treatment initiation can potentially happen in a wide time range and since the length of follow-up required to observe the event of interest may not be available, especially for untreated patients, analysis methods should use the available data to the best extent possible while producing interpretable, actionable results. A clinical and statistical question of interest centers on whether to quantify treatment differences using the time since birth or the time since treatment initiation. Motivated by the question arising in the analysis of time to loss of ambulation, a key disease progression milestone for Duchenne Muscular Dystrophy patients, we assess the statistical impact of choosing between the two time axes in various scenarios that could be observed in controlled trials or a real world setting using a comprehensive simulation study. We compare the performance of the two approaches in terms of their ability to optimally use the information available in the data. We also discuss the interpretability and causal implications of results from the two approaches using a comparison of treated patients with external controls as an example. Further, we discuss the problem of

time zero or index time definition in real-world settings and provide recommendations to avoid immortal time bias.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S58: OPPORTUNITIES AND CHALLENGES IN OMICS DATA ANALYSIS AND INTEGRATION

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Southern Ground A/B (Mezzanine Level)

Organizer: Shilin Zhao, Vanderbilt University Medical Center
Chair: Shilin Zhao, Vanderbilt University Medical Center

3:00 PM–3:20 PM Speaker: Xiao Dong, University of Minnesota

Single-cell analysis of somatic DNA mutations in normal, noncancerous tissues

Author(s): Xiao Dong, University of Minnesota

Single-cell sequencing for analyzing DNA mutations across the genome in somatic tissues is critically important for studying development, cancer, and aging. However, current procedures are prone to artifacts and a reliable protocol for single-cell somatic mutation analysis remains to be improved. With our newly developed methods in single-cell whole-genome sequencing, we analyzed multiple types of primary cells from humans and mice varying in age and smoking status. The results strongly suggest that spontaneous somatic mutations accumulating with age reach high enough levels to contribute to age-related functional decline, such as the well-documented cell-intrinsic changes. Taken together, our single-cell sequencing method provides a firm foundation for analyzing cellular genetic heterogeneity in normal, noncancerous tissues.

3:20 PM–3:20 PM Speaker: Shilin Zhao, Vanderbilt University Medical Center

stImage: A versatile framework for optimizing spatial transcriptomic analysis through customizable deep histology and location informed integration

Author(s): Yu Wang, Vanderbilt University; Haichun Yang, Vanderbilt University; Ruining Deng, Vanderbilt University; Yuankai Huo, Vanderbilt University; Qi Liu, Vanderbilt University; Yu Shyr, Vanderbilt University; Shilin Zhao, Vanderbilt University

Spatial transcriptomics maps out organizational structures of cells with their genome-wide transcriptional profiles and histology images. Fully exploiting these three different views of data holds great promise to characterize spatial expression heterogeneity accurately. Although several methods have been developed to perform location- and/or histology-informed integration of spatial transcriptomics, they are not one-size-fits-all approaches, but only perform well for certain data and conditions. Here, we present stImage, a comprehensive and flexible framework for optimizing spatial transcriptomic analysis. stImage provides 54 integrative strategies and allows users to develop customized pipelines freely. We illustrate the benefits of stImage for spatial cell clustering on a variety of datasets, demonstrating its superior performance by selecting optimal integrative strategies in different datasets.

3:40 PM–4:00 PM Speaker: Tianyuan Yao, Vanderbilt University

Spatial pathomics toolkit for quantitative analysis of podocyte nuclei with histology and spatial transcriptomics data in renal pathology

Author(s): Jiayuan Chen, Vanderbilt University; Yu Wang, Vanderbilt University Medical Center; Ruining Deng, Vanderbilt University; Can Cui, Vanderbilt University; Tianyuan Yao, Vanderbilt University; Jianyong Zhong, Vanderbilt University Medical Center; Agnes Fogo, Vanderbilt University; Haichun Yang, Vanderbilt University Medical Center; Shilin Zhao, Vanderbilt University Medical Center; Yuankai Huo, Vanderbilt University

Podocytes, specialized epithelial cells that envelop the glomerular capillaries, play a pivotal role in maintaining renal health. The current description and quantification of features on pathology slides are limited, prompting the need for innovative solutions to comprehensively assess diverse phenotypic attributes within Whole Slide Images (WSIs). In particular, understanding the morphological characteristics of podocytes, terminally differentiated glomerular epithelial cells, is crucial for studying glomerular injury. This paper introduces the Spatial Pathomics Toolkit (SPT) and applies it to podocyte pathomics. The SPT consists of three main components: (1) instance object segmentation, enabling precise identification of podocyte nuclei; (2) pathomics feature generation, extracting a comprehensive array of quantitative features from the identified nuclei; and (3) robust statistical analyses, facilitating a comprehensive exploration of spatial relationships between morphological and spatial transcriptomics features. The SPT successfully extracted and analyzed morphological and textural features from podocyte nuclei, revealing a multitude of podocyte morphomic features through statistical analysis. Additionally, we demonstrated the SPT's ability to unravel spatial information inherent to podocyte distribution, shedding light on spatial patterns associated with glomerular injury. By disseminating the SPT, our goal is to provide the research community with a powerful and user-friendly resource that advances cellular spatial pathomics in renal pathology.

4:00 PM–4:30 PM **Q&A and Floor Discussion**

S59: KISS SESSION II: RECENT ADVANCES IN STATISTICAL METHODS FOR COMPLEX DATA

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Blackbird B (Mezzanine Level)

Organizer: Yeonhee Park, University of Wisconsin-Madison
Chair: Yeonhee Park, University of Wisconsin-Madison
KISS = Korean International Statistical Society

3:00 PM–3:20 PM Speaker: Hyo Young Choi, University of Tennessee Health Science Center

ArtistR: A novel framework for systematically detecting alternative transcript initiation by integrating ATAC-seq and RNA-seq

Author(s): Hyo Young Choi, University of Tennessee Health Science Center; Won-Young Choi, University of Tennessee Health Science Center; D. Neil Hayes, University of Tennessee Health Science Center

Alternative transcription initiation (ATI) has been frequently observed in cancer suggesting that it contributes to the malignant transformation of the cells. However, ATI remains largely unexplored mainly due to the lack of tools for detecting ATI. We propose a multi-omics integration method that integrates base-level RNA-seq and ATAC-seq, which aims to identify ATIs as well as characterize the latent structure of the underlying isoforms. This method equips many unique features that can be useful for the identification of novel ATIs. Because the method scans an entire collection of DNA accessible regions provided by ATAC-seq, it enables a comprehensive screening of novel ATI candidates that are not limited to the known promoters. Additionally, the uncompressed view of base-resolution RNA-seq coverage allows us to infer the structure of individual isoforms independently of known gene annotation. We applied this method to a sub-cohort (N=350) of TCGA pan-cancer samples and successfully identified the ATIs including some challenging cases such as ATIs located at internal introns or at constitutive exons in other transcripts.

3:20 PM–3:20 PM Speaker: Jaihee Choi, Rice University

Bayesian variable selection for interval-censored outcomes in genome-wide association studies
Author(s): Jaihee Choi, Rice University; Ryan Sun, University of Texas MD Anderson Cancer Center

With the growing popularity of genetic and health databases such as the UK Biobank, there is increased access to Genome-wide association studies (GWAS) with interval-censored time-to-event outcomes. Gene set-based association tests have proven to be successful in identifying genes or risk loci associated with outcomes of interest while maintaining sufficient statistical power. However, fine-mapping the specific SNP or SNPs within these gene sets associated with the disease can lead to better understanding of the genetic etiology of the disease. Though using interval-censored time-to-event outcomes can provide more information about the genetic pathology behind disease more than the binary or right-censored representation of the data, there currently are not many methods that work with interval-censored outcomes. In this work, we investigate a Bayesian framework for fine-mapping individual genetic variants associated with interval-censored data. We apply this framework to colorectal cancer data from the UK Biobank.

3:40 PM–4:00 PM Speaker: Ray Bai, University of South Carolina

Generative quantile regression with variability penalty

Author(s): Shijie Wang, University of South Carolina; Minsuk Shin, Gauss Lab; Ray Bai, University of South Carolina

Quantile regression and conditional density estimation can reveal structure that is missed by mean regression, such as multimodality and skewness. In this talk, we introduce a deep learning generative model for joint quantile estimation called Penalized Generative Quantile Regression (PGQR). Our approach simultaneously generates samples from many random quantile levels, allowing us to infer the conditional distribution of a response variable given a set of covariates. Our method employs a novel variability penalty to avoid the problem of vanishing variability, or memorization, in deep generative models. Further, we introduce a new family of partial monotonic neural networks (PMNN) to circumvent the problem of crossing quantile curves. A major benefit of PGQR is that it can be fit using a single optimization, thus bypassing the need to repeatedly train the model at multiple quantile levels or use computationally expensive cross-validation to tune the penalty parameter. We illustrate the efficacy of PGQR through extensive simulation studies and analysis of real datasets.

4:00 PM–4:20 PM Speaker: Yunxiang Huang, University of California, San Francisco

A synthetic data approach for incorporating an external individual level prediction information in the proportional hazards model

Author(s): Yunxiang Huang, University of California, San Francisco; Yena Jeon, University of California, San Francisco; Mi-Ok Kim, University of California, San Francisco

Individualized risk prediction algorithms, such as the Prostate Cancer Risk Assessment tool, are increasingly utilized for the risk prediction of cancer relapse or progression. Since these algorithms are commonly trained on large datasets, it is reasonable to expect that effectively integrating information from these trained prediction algorithms into the inference procedure can improve the efficiency of analyzing individual studies. In this research, we consider Cox regression analysis of right-censored time-to-event outcome by incorporating external information generated by these large-scale prediction model. We take a synthetic data approach to generate pseudo-observations corresponding to the external prediction information and consider the full likelihood for the combined data. We establish the existence and uniqueness of the maximum combined likelihood estimator. An expectation-maximization algorithm is proposed to effectively obtain this estimator, with guarantees of global optimality. The performance of the proposed model is illustrated through simulations and an application to prostate cancer trial data.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S60: INTEGRATING INFORMATION FROM DIFFERENT DATA SOURCES: SOME NEW DEVELOPMENTS

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Gold (Lower Level)

Organizer and Chair: Peisong Han, Gilead Sciences, Inc.

3:00 PM–3:20 PM Speaker: Yajuan Si, University of Michigan

On the use of auxiliary variables in multilevel regression and poststratification

Author(s): Yajuan Si, University of Michigan

Multilevel regression and poststratification (MRP) is a popular method for addressing selection bias in subgroup estimation, with broad applications across fields from social sciences to public health. In this paper, we examine the inferential validity of MRP in finite populations, exploring the impact of poststratification and model specification. The success of MRP relies heavily on the availability of auxiliary information that is strongly related to the outcome. To enhance the fitting performance of the outcome model, we recommend modeling the inclusion probabilities conditionally on auxiliary variables and incorporating flexible functions of estimated inclusion probabilities as predictors in the mean structure. We present a statistical data integration framework that offers robust inferences for probability and nonprobability surveys, addressing various challenges in practical applications. Our simulation studies indicate the statistical validity of MRP, which involves a tradeoff between bias and variance, with greater benefits for subgroup estimates with small sample sizes, compared to alternative methods. We have applied our methods to the Adolescent Brain Cognitive Development (ABCD) Study, which collected information on children across 21 geographic locations in the U.S. to provide national representation, but is subject to selection bias as a nonprobability sample. We focus on the cognition measure of diverse groups of children in the ABCD study and show that the use of auxiliary variables affects the findings on cognitive performance.

3:20 PM–3:20 PM Speaker: Lu Tang, University of Pittsburgh

Learning of robust individualized decision rules in heterogeneously distributed data

Author(s): Lu Tang, University of Pittsburgh

This work introduces a learning algorithm for deriving individualized decision rules (IDRs) that are robust to distributional uncertainty in heterogeneous data sources. It is motivated by the need to uniformly improve decision-making across multiple hospitals of a single health system. Traditional approaches assume that data are sampled from a single population of interest. With multiple hospitals that vary in patient populations and treatments, an IDR that is effective in one hospital may not be as effective in another. Due to distributional heterogeneity, the performance achieved by a globally optimal IDR varies greatly across sites, preventing it from being safely applied to unseen sites. Additionally, data from multiple hospitals cannot be pooled due to privacy concerns. To address these challenges, we developed an federated learning framework to learn IDRs from distributed data. The proposed framework introduces a conditional maximin objective to enhance individual outcomes across sites, ensuring robustness against site variations. The proposed method effectively improves the generalizability of IDRs for the management of sepsis in a hospital network.

3:40 PM–4:00 PM Speaker: Dungang Liu, University of Cincinnati

Surrogate method for partial association between mixed data with application to well-being survey analysis

Author(s): Shaobo Li, University of Kansas; Zhaohu Fan, Georgia Institute of Technology; Ivy Liu, University of Victoria at Wellington; Philip Morrison, University of Victoria at Wellington; Dungang Liu, University of Cincinnati

This paper is motivated by the analysis of a survey study focusing on college student well-being before and after the COVID-19 pandemic outbreak. A statistical challenge in well-being studies lies in the multidimensionality of outcome variables, recorded in various scales such as continuous, binary, or ordinal. The presence of mixed data complicates the examination of their relationships when adjusting for important covariates. To address this challenge, we propose a unifying framework for studying partial association between mixed data. We achieve this by defining a unified residual using the surrogate method. The idea is to map the residual randomness to a consistent continuous scale, regardless of the original scales of outcome variables. This framework applies to parametric or semiparametric models for covariate adjustments. We validate the use of such residuals for assessing partial association, introducing a measure that generalizes classical Kendall's tau to capture both partial and marginal associations. Moreover, our development advances the theory of the surrogate method by demonstrating its applicability without requiring outcome variables to have a latent variable structure. In the analysis of the college student well-being survey, our proposed method unveils the contingency of relationships between multidimensional well-being measures and micro personal risk factors (e.g., physical health, loneliness, and accommodation), as well as the macro disruption caused by COVID-19.

4:00 PM–4:20 PM Speaker: Sixia Chen, University of Oklahoma Health Sciences Center

Combining probability and non-probability samples using semi-parametric quantile regression and a non-parametric estimator of the participation probability

Author(s): Sixia Chen, University of Oklahoma Health Sciences Center; Emily Berg, Iowa State University; Cindy Yu, Iowa State University

Non-probability samples are prevalent in various fields, such as biomedical studies, educational research, and business investigations, owing to the escalating challenges associated with declining response rates and the cost-effectiveness and convenience of utilizing such samples. However, relying on naive estimates derived from non-probability samples, without adequate adjustments, may introduce bias into study outcomes. Addressing this concern, data integration methodologies, which amalgamate information from both probability and non-probability samples, have demonstrated effectiveness in mitigating selection bias. Commonly employed data integration approaches encompass mass imputation, propensity score weighting, and hybrid methodologies. Nonetheless, the efficacy of these methods hinges upon the assumptions underlying the models. This paper introduces innovative and robust data integration approaches, notably a semi-parametric quantile regression-based mass imputation approach and a doubly robust approach that integrates a non-parametric estimator of the participation probability for non-probability samples. Our proposed methodologies exhibit greater robustness compared to existing parametric approaches, particularly concerning model misspecification and outliers. Theoretical results are established, including variance estimators for our proposed estimators.

Through comprehensive simulation studies and real-world applications, our findings demonstrate the promising performance of the proposed estimators in reducing selection bias and facilitating valid statistical inference. This research contributes to the advancement of robust methodologies for handling non-probability samples, thereby enhancing the reliability and validity of research outcomes across diverse domains.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S61: STATISTICAL INNOVATIONS FOR UNDERSTANDING HUMAN MICROBIOME AND THEIR INTERACTIONS WITH OTHER OMICS

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Melody (Lobby Level)

Organizer: Wodan Ling, Weill Cornell Medicine
Chair: Yuan Zhang, University of Pennsylvania

3:00 PM–3:20 PM Speaker: Amarise Little, Fred Hutch Cancer Center

Visualizing longitudinal microbiome data

Author(s): Amarise Little, Fred Hutch Cancer Center; Michael Wu, Fred Hutch Cancer Center

Principal Coordinate Analysis (PCoA) is a cornerstone method for visualizing microbiome data, but its application to longitudinal studies remains uncertain. In this project, we address this gap by proposing a framework for visualizing longitudinal microbiome data through Adjusted Kernel Principal Component Analysis. Building upon the concept of covariate-adjusted PCoA, our approach embeds microbiome data into kernel matrices to capture multivariate profiles while addressing the complexities of repeated measurements. We demonstrate the effectiveness of our method in distilling essential axes of microbiome community variation while providing nuanced insights into temporal dynamics. By using residuals from linear mixed models, we overcome challenges associated with dependencies and enhance visualizations. Our framework offers a robust approach for uncovering temporal patterns in microbial communities, thereby facilitating informed decision-making in microbiome research.

3:20 PM–3:20 PM Speaker: Tianying Wang, Colorado State University

A high-dimensional calibration method for log-contrast models subject to measurement errors

Author(s): Huali Zhao, Tsinghua University; Tianying Wang, Colorado State University

Motivated by the challenges in analyzing gut microbiome and metagenomic data, this work focuses on addressing measurement errors in high-dimensional regression models that involve compositional covariates. As the initial effort to conduct statistical inference on high-dimensional compositional data affected by measurement inaccuracies, a calibration method is introduced for the linear log-contrast model. Under certain mild conditions regarding the sparsity level of the parameter, we established the asymptotic normality of the estimator for inference. Through numerical experiments, it has been shown that our high-dimensional calibration approach can effectively reduce the bias and achieve the nominal coverage rate for the confidence intervals. We demonstrate our method by a microbiome study that explores the relationship between human body mass index and gut microbiome composition. Furthermore, the applicability of our proposed method extends beyond compositional data, offering the potential for broader applications in various research studies.

3:40 PM–4:00 PM Speaker: Siyuan Ma, Vanderbilt University Medical Center

CompDA: Defining and finding a new type of health-microbiome associations

Author(s): Siyuan Ma, Vanderbilt University Medical Center; Curtis Huttenhower, Harvard T.H. Chan School of Public Health; Lucas Janson, Harvard University

A major task of microbiome epidemiology is association analysis, where the goal is to iden-

tify microbial features related to host health. This is commonly performed by the differential abundance (DA) analysis, which, by design, examines each microbe as isolated from the rest of the microbiome. This does not properly account for the microbiome's compositional nature or microbe-microbe ecological interactions, and can lead to confounded findings, i.e., microbes that only appear to associate with health through their confounding association with health-related, biologically informative microbes. To remedy these issues, we present Compositional Differential Abundance (CompDA) analysis, a novel approach for health-microbiome association. CompDA provides a novel approach to identify health-related microbes by examining the microbiome holistically, which a) accounts for the data's compositionality and ecological interactions, and b) has clear interpretations corresponding to host health as affected by microbiome-based interventions. CompDA prioritizes health-related microbes and controls false discoveries by implementing recent advances from high-dimensional statistics, and can be flexibly adapted to many common tasks in modern microbiome epidemiology, including enhancing microbiome-based machine learning by providing rigorous p-values to prioritize important features. We validate the performance of CompDA, and compare against canonical microbiome association methods including DA with extensive, real-data-informed simulation studies. Lastly, we report novel and consistent findings of CompDA in application studies, including a) recently reported microbial signatures of colorectal cancer from cross-study machine learning, and b) well-established microbial associations of early onset Crohn's disease in a pediatric cohort.

4:00 PM–4:20 PM Speaker: Wodan Ling, Weill Cornell Medicine

Association analysis for microbiome biomarkers via zero-inflated quantile-based approaches
Author(s): Chenlian Fu, Weill Cornell Medicine; Jiuyao Lu, Johns Hopkins University; Ni Zhao, Johns Hopkins University; Wodan Ling, Weill Cornell Medicine

Understanding the role of the human microbiome in health and disease is crucial to elucidate the underlying biomedical mechanism or devise new intervention paradigms. However, as microbiome data is sparse and over-dispersed, tailed methods often fail to control type I error due to unsatisfied distributional assumptions, while classical methods suffer from loss of power as they barely take full advantage of the data characteristics. We will present a series of zero-inflated quantile-based approaches for microbiome association analysis, which use logistic-type regression to assess the clinical variable's effect on the presence-absence status of the investigated taxon, and a quantile rank-score test or process adjusted for zero inflation to evaluate the variable's effect on the taxon abundance distribution given its presence. The approaches are robust to irregular microbiome distributions and allow for the assessment of the variable's effect on the overall distribution of the taxon, enabling the detection of heterogeneous associations. Application of the methods to real microbiome studies demonstrates their improved robustness and power compared to existing methods.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S62: RECENT ADVANCES IN TIME SERIES ANALYSIS FOR BUSINESS AND ECONOMICS DATA

Tuesday, June 18, 2024
3:00 PM–4:30 PM, Green Room (Lobby Level)

Organizer and Chair: Chun Yip Yau, Chinese University of Hong Kong

3:00 PM–3:20 PM Speaker: Chun Yip Yau, Chinese University of Hong Kong

Burn-in selection in simulating time series

Author(s): Chun Yip Yau, Chinese University of Hong Kong

Many time series models are defined in a recursive manner, which prohibits exact simulations. In practice, one appeals to simulating a long time series and discarding a large number of initial simulated observations, known as the burn-in. For autoregressive models where the dependence decays exponentially fast, the choice of the burn-in is not critical. However, for long-memory time series where the dependence from the remote past is strong, it is not clear how to select the burn-in number. By combining several samplers with randomized burn-in numbers, we develop a method for exactly simulating the expectation of a statistic computed from a time series. Moreover, with some suitably chosen statistics, the exact simulation method can be applied to quantify the effect of burn-in numbers on the simulated sample. Simulation studies are conducted to provide some practical guidance for burn-in selections.

3:20 PM–3:20 PM Speaker: Zifeng Zhao, University of Notre Dame

High-dimensional dynamic pricing under non-stationarity: Learning and earning with change-point detection

Author(s): Zifeng Zhao, Feiyu Jiang, Yi Yu, Xi Chen

We consider a high-dimensional dynamic pricing problem under non-stationarity, where a firm sells products to T sequentially arriving consumers that behave according to an unknown demand model with potential changes at unknown times. The demand model is assumed to be a high-dimensional generalized linear model (GLM), allowing for a feature vector in $\mathbb{R}^d$ that encodes products and consumer information. To achieve optimal revenue (i.e., least regret), the firm needs to learn and exploit the unknown GLMs while monitoring for potential change-points. To tackle such a problem, we first design a novel penalized likelihood-based online change-point detection algorithm for high-dimensional GLMs, which is the first algorithm in the change-point literature that achieves optimal minimax localization error rate for high-dimensional GLMs. A change-point detection assisted dynamic pricing (CPDP) policy is further proposed and achieves a near-optimal regret of order $O(s\sqrt{\Upsilon_T T}\log(Td))$, where s is the sparsity level, and Υ_T is the number of change-points. This regret is accompanied with a minimax lower bound, demonstrating the optimality of CPDP (up to logarithmic factors). In particular, the optimality with respect to Υ_T is seen for the first time in the dynamic pricing literature and is achieved via a novel accelerated exploration mechanism. Extensive simulation experiments and a real data application on online lending illustrate the efficiency of the proposed policy and the importance and practical value of handling non-stationarity in dynamic pricing.

3:40 PM–4:00 PM Speaker: Wai Leong Ng, Hang Seng University of Hong Kong

Inference for multiple change-points in piecewise locally stationary time series
Author(s): Wai Leong Ng, Hang Seng University of Hong Kong

Change-point detection and locally stationary time series modeling are two major approaches for the analysis of non-stationary time series. The former aims to identify abrupt changes in the parameters of a stationary time series model, while the latter employs time-varying parameter curves to describe smooth changes in the mean or dependence structure of a time series. However, in some applications, abrupt and smooth changes can co-exist, and neither of the two approaches alone can model the data adequately. In this paper, we propose a likelihood based procedure for the inference of multiple change-points in a locally stationary time series. In contrast to traditional change-point analysis where an abrupt change occurs in a real-valued parameter, a change in locally stationary time series occurs in a parameter curve, and can be classified as a jump or a kink depending on whether the curve is discontinuous or not. We show that the proposed procedure can consistently estimate the number, locations, and the types of change-points. Two different asymptotic distributions corresponding respectively to jump and kink estimators are also established. Extensive simulation studies and a real data application to financial time series are provided.

4:00 PM–4:20 PM Speaker: Haihan Yu, University of Rhode Island

A blockwise empirical likelihood method for time series in frequency domain inference
Author(s): Haihan Yu, University of Rhode Island; Mark Kaiser, Iowa State University; Daniel Nordman, Iowa State University

Frequency domain analysis of time series is often difficult, as periodogram-based statistics involve non-linear averages with complicated variances. Due to the latter, nonparametric approximations from resampling or empirical likelihood (EL) are useful. However, current versions of periodogram-based EL for time series are highly restricted: these are valid only for linear processes and for special parameters (i.e., ratios). For general frequency domain inference with stationary, weakly dependent time series, we develop a spectral EL (SEL) method by combining two previously separate EL frameworks for time series: block-based EL and periodogram-based EL. This hybridization strategy is new and theoretically non-trivial, particularly as existing block-based EL relies on time domain averages that differ substantially from frequency domain counterparts. We formulate SEL statistics for parameters based on spectral estimating functions and periodogram subsamples. Under mild conditions, SEL log-ratio statistics are shown to be well-defined, admitting chi-square limits. Further, we formally establish an effective bootstrap procedure coupled with SEL. As a result, the SEL method can be used for nonparametric, asymptotically correct confidence regions and tests for frequency domain inference without explicit estimation of intricate variances of periodogram-based statistics. This broadly extends the applicability of EL for time series in three directions: (i) SEL can treat any spectral mean parameters; (ii) SEL is valid for both linear and non-linear processes; and (iii) SEL has a provable bootstrap development, which is rare for time series EL, and provides a novel alternative to other resampling approximations in the frequency domain. Simulation studies suggest the proposed method performs well compared to other non-EL approaches. A real data example demonstrates that SEL has application and extension to complicated scenarios.

4:20 PM–4:30 PM **Q&A and Floor Discussion**

S63: BIOSTATISTICS IN GOVERNMENT AND PHARMACEUTICAL INDUSTRY

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Blackbird A (Mezzanine Level)

Organizer: Panpan Zhang, Vanderbilt University Medical Center
Chair: Panpan Zhang, Vanderbilt University Medical Center

5:00 PM–5:20 PM Speaker: Li Cheung, National Cancer Institute, NIH

Risk and benefit models for cancer screening guidelines

Author(s): Li Cheung, National Cancer Institute, NIH

I'll be speaking of novel risk and life-gained benefit models for cancer screening guidelines, using examples from my work on lung, cervical, and oral cancer screening. The talk will briefly touch upon three concepts: (1) moving from risk-based to life-gained benefit-based precision screening with application to determining lung cancer screening eligibility, (2) prevalent and incident risk models for preclinical outcomes with application to cervical precancer detection, and (3) multistate models to inform the clinical management of oral precancer lesions that can disappear, reappear, or progress to cancer.

5:20 PM–5:20 PM Speaker: Phoebe Jiang, Biogen

Recent advances in comparative effectiveness and precision medicine

Author(s): Phoebe Jiang, Biogen

This presentation advocates for advancing comparative effectiveness research beyond average treatment effects by integrating precision medicine (PM), causal inference, and machine learning (ML). We first explore recent breakthroughs in biomarker discovery, subgroup analysis, and interpretable techniques. A proposed pipeline that combines PM principles with ML to address baseline imbalances and identify heterogeneous treatment effects is then introduced. Through a real-world study on relapsing-remitting multiple sclerosis patients, we demonstrate the application of this approach in assessing treatment efficacy at both aggregate and individual levels.

5:40 PM–6:00 PM Speaker: Cong Wang, US Food and Drug Administration

CBER's statistical review experience with CAR-T products

Author(s): Cong Wang, US Food and Drug Administration

On April 5, 2024, the FDA approved the CAR-T product CARVYKTI for the treatment of adult patients with relapsed or refractory multiple myeloma, who have received at least 1 prior line of therapy, including a proteasome inhibitor, and an immunomodulatory agent, and are refractory to lenalidomide. At the time of this efficacy supplement submission with data cutoff date of November 1, 2022, the applicant provided the results of an interim analysis of OS with 34% information fraction. At this interim OS analysis, the OS Kaplan-Meier curves crossed at approximately 10 months, with inferior OS in the CARVYKTI arm compared to the standard

of care arm prior to 10 months post randomization. Although there was observed early OS detriment, this concern appears mitigated by the subsequent long-term benefits. Given the collective statistical evidence, including clinically meaningful improvements in the primary and key secondary endpoints PFS, CRR and ORR in a difficult-to-treat patient population and life-threatening nature of the disease, FDA statistical review recommended approval for the applicant's proposed indication. However, caution is warranted regarding the OS results, and longer follow-up is necessary to further confirm the long-term OS benefit.

6:00 PM–6:20 PM Speaker: Lanju Zhang, Vertex Pharmaceuticals

Statistics in drug development: an industry perspective
Author(s): Lanju Zhang, Vertex Pharmaceuticals

In the data driven drug development process, statistics plays an indispensable role in study design and analysis to ensure answering the right questions, collecting right data, running right analysis, and providing right interpretation and reporting of the study results. In this talk, we will introduce different statistical methods used in different stages of drug development process and the role and qualifications of statisticians in developing and executing these statistical methods. Personal experiences will be shared.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

S64: EXPLORING THE IMPACT OF HIGH-DIMENSIONAL ENVIRONMENTAL MIXTURE ON HEALTH OUTCOME

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Lyric (Lobby Level)

Organizer: Zhen Chen, National Institute of Child Health and Human Development (NICHD), NIH
Chair: Danping Liu, National Cancer Institute, NIH

5:00 PM–5:20 PM Speaker: Shelley Liu, Icahn School of Medicine at Mount Sinai

Mixture item response theory to quantify cumulative environmental exposure burden

Author(s): Shelley Liu, Icahn School of Medicine at Mount Sinai; Leah Feuerstahler, Fordham University; Yitong Chen, Icahn School of Medicine at Mount Sinai; Joseph Braun, Brown University; Jessie Buckley, University of North Carolina at Chapel Hill

Quantifying a person's cumulative exposure burden to environmental toxicants is important for risk assessment. However, different people may be exposed to different sets of environmental toxicants due to heterogeneity in exposure sources and patterns. We used mixture item response theory to estimate a person's total exposure burden to per- and polyfluoroalkyl substances (PFAS), dubbed toxic forever chemicals, while accounting for the fact that different people have different diets and behaviors that may expose them to different sets of PFAS chemicals. This ensures that PFAS burden scores can be equitably compared across population subgroups. We applied our methods to PFAS biomonitoring data from the National Health and Nutrition Examination Survey (NHANES) 2013-2018, where we found that Asian Americans have significantly higher PFAS burden compared with non-Hispanic Whites, but this disparity was masked when using summed PFAS concentrations as the exposure metric. Our work suggests that risk assessment may want to consider a summary exposure metric for environmental toxicants that accounts for exposure heterogeneity, so that the summary metric used is fair and informative for all people.

5:20 PM–5:20 PM Speaker: Debamita Kundu, University of Virginia

Bayesian inference of chemical mixtures in risk assessment incorporating the hierarchical principle

Author(s): Debamita Kundu, University of Virginia; Sungduk Kim, National Institutes of Health; Paul S. Albert, National Institutes of Health

Analyzing health effects associated with exposure to environmental chemical mixtures is a challenging problem in epidemiology, toxicology, and exposure science. In particular, when there are a large number of chemicals under consideration it is difficult to estimate the interactive effects without incorporating reasonable prior information. Based on substantive considerations, researchers believe that true interactions between chemicals need to incorporate their corresponding main effects. In this paper, we use this prior knowledge through a shrinkage prior that a priori assumes an interaction term can only occur when the corresponding main effects exist. Our initial development is for logistic regression with linear chemical effects. We extend this formulation to include non-linear exposure effects and to account for exposure subject to detection limit. We develop an MCMC algorithm using a shrinkage prior that shrinks the interaction terms closer to zero as the main effects get closer to zero. We examine the performance of our methodology through simulation studies and illustrate an analysis of chemical interactions in a case-control study in cancer.

5:40 PM–6:00 PM Speaker: Madeleine St. Ville, National Institute of Child Health and Human Development (NICHD), NIH

A Bayesian variable selection approach for identifying the latency period between gestational weight gain and fetal development

Author(s): Madeleine St. Ville, National Institute of Child Health and Human Development (NICHD), NIH; Zhen Chen, NICHD; Katherine Grantz, NICHD

Inadequate or excessive gestational weight gain (GWG) during pregnancy is associated with an increased risk of pregnancy complications, such as abnormal fetal growth. The timing of GWG may modify its impact on fetal development, and the precise identification of a latency period could aid in clinical decision-making. Latency selection procedures based on effect sizes have been shown to be biased, while those relying on model goodness of fit tend to underestimate variability. Other work has resorted to approximations to tackle the challenging problem of estimating latency, a discrete quantity, or to estimation methods that may lack theoretical justifications. We take a Bayesian variable selection perspective in the framework of distributed lag models and jointly estimate the latency parameter and the associated effect size between GWG and a fetal development measure. With a sample from the posterior distribution via Markov chain Monte Carlo, adequate measures of variability of the latency parameter can be obtained. In addition, the effect size estimate represents a weighted average across different latency scenarios. We evaluate the performance of this method via extensive simulation studies. Then our proposed approach is applied to an analysis of the relationship between GWG from conception to 37 gestational weeks, and fetal fractional arm volume at 37 weeks using data from the National Institute of Child Health and Human Development (NICHD) Fetal 3D Study, a prospective cohort of women with singleton pregnancies.

6:00 PM–6:20 PM Speaker: Ruijin Lu, Washington University in St. Louis

Bayesian single index models with additive regression trees

Author(s): Ruijin Lu, Zhen Chen

In analyzing data of environmental mixtures, single (SIM) and multiple index models (MIM) are powerful tools given their nonparametric links and interpretable index coefficients. In this paper, we take a Bayesian additive regression tree (BART) perspective to these models and consider variable selections, particularly the selection of exposures in the indices. The challenge of applying BART to SIM/MIM is tackled by using a sigmoid gating function in place of the binary routine at each splitting node and that of variable selection by placing a sparsity inducing Dirichlet hyperprior. We examine the performance of the proposed approach by conducting extensive simulations and apply it to commonly used benchmark data sets. For real data application, we investigate the link between birth weight and exposures to environmental pollutants, dietary intakes, and physical activities during pregnancy using data from the National Institute of Child Health and Human Development (NICHD) Fetal Growth Study.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

S65: SOME RECENT ADVANCES IN SPATIAL DATA

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Ocean Way (Mezzanine Level)

Organizer: Didong Li, University of North Carolina at Chapel Hill
Chair: Didong Li, University of North Carolina at Chapel Hill

5:00 PM–5:20 PM Speaker: Luhuan Wu, Columbia University

Practical and asymptotically exact conditional sampling in diffusion models

Author(s): Luhuan Wu, Columbia University; Brian Trippe, Columbia University; Christian Naesseth, University of Amsterdam; David Blei, Columbia University; John Cunningham, Columbia University

Diffusion models have been successful on a range of conditional generation tasks including molecular design and text-to-image generation. However, these achievements have primarily depended on task-specific conditional training or error-prone heuristic approximations. Ideally, a conditional generation method should provide exact samples for a broad range of conditional distributions without requiring task-specific training. To this end, we introduce the Twisted Diffusion Sampler, or TDS. TDS is a sequential Monte Carlo (SMC) algorithm that targets the conditional distributions of diffusion models. The main idea is to use twisting, an SMC technique that enjoys good computational efficiency, to incorporate heuristic approximations without compromising asymptotic exactness. We first find in simulation and on MNIST image inpainting and class-conditional generation tasks that TDS provides a computational statistical trade-off, yielding more accurate approximations with many particles but with empirical improvements over heuristics with as few as two particles. We then turn to motif-scaffolding, a core task in protein design, using a TDS extension to Riemannian diffusion models. On benchmark test cases, TDS allows flexible conditioning criteria and often outperforms the state of the art.

5:20 PM–5:20 PM Speaker: Jian Hu, Emory University

MorphLink: Bridging cellular morphological behaviors and molecular dynamics in multi-modal spatial omics

Author(s): Jing Huang, Emory University; Linghua Wang, University of Texas MD Anderson Cancer Center; Jian Hu, Emory University

The increasing generation of multi-modal spatial omics data is invaluable for exploring aberrant cellular behavior in diseases from both morphological and molecular perspectives. Current analytical methods primarily focus on predictive tasks, such as clustering, and do not adequately examine the relationship between cell morphology and molecular dynamics. Here, we present MorphLink, a framework designed to systematically identify disease-related morphological-molecular interplays. MorphLink has been evaluated across a wide array of datasets, showcasing its effectiveness in extracting and linking interpretable morphological features with various molecular measurements in bi- and tri-modality spatial omics analyses. These linkages provide a transparent depiction of cellular behaviors that drive tumor heterogeneity and immune diversity across different cancers. Additionally, MorphLink is scalable and robust against cross-sample batch effects. We envision MorphLink becoming a significant tool for multi-sample, multi-modality spatial omics data analysis, facilitating atlas usage, and enhancing the interpretation of analyses.

5:40 PM–6:00 PM Speaker: Aritra Halder, Drexel University

Bayesian modeling with spatial curvature processes

Author(s): Aritra Halder, Drexel University; Sudipto Banerjee, University of California at Los Angeles; Dipak K. Dey, University of Connecticut

Spatial process models are widely used for modeling point-referenced variables arising from diverse scientific domains. Analyzing the resulting random surface provides deeper insights into the nature of latent dependence within the studied response. We develop Bayesian modeling and inference for rapid changes on the response surface to assess directional curvature along a given trajectory. Such trajectories or curves of rapid change, often referred to as wombling boundaries, occur in geographic space in the form of rivers in a flood plain, roads, mountains or plateaus or other topographic features leading to high gradients on the response surface. We demonstrate fully model based Bayesian inference on directional curvature processes to analyze differential behavior in responses along wombling boundaries. We illustrate our methodology with a number of simulated experiments followed by multiple applications featuring the Boston Housing data; Meuse river data; and temperature data from the Northeastern United States. Supplementary materials for this talk are available online.

6:00 PM–6:20 PM Speaker: Didong Li, University of North Carolina at Chapel Hill

Investigating spatial dynamics in spatial omics data

Author(s): Jiawen Chen, University of North Carolina at Chapel Hill; Caiwei Xiong, University of North Carolina at Chapel Hill; Quan Sun, University of North Carolina at Chapel Hill; Geoffery Wang, North Carolina State University; Gaorav Gupta, University of North Carolina at Chapel Hill; Aritra Halder, Drexel University; Yun Li, University of North Carolina at Chapel Hill; Didong Li, University of North Carolina at Chapel Hill

Spatial omics technologies revolutionize our view of biological processes within tissues. However, existing methods fail to capture localized, sharp changes characteristic of critical events (e.g., tumor development). Here, we present StarTrail, a novel gradient based method that powerfully defines rapidly changing regions and detects "cliff genes"—genes exhibiting drastic expression changes at highly localized or disjoint boundaries. StarTrail, the first to leverage spatial gradients for spatial omics data, also quantifies directional dynamics. Across multiple datasets, StarTrail accurately delineates boundaries (e.g., brain layers, tumor-immune boundaries), and detects cliff genes that may regulate molecular crosstalk at these biologically relevant boundaries but are missed by existing methods. For instance, StarTrail precisely pinpointed the cancer-immune interface in a HER2+ breast cancer dataset, unveiled key cliff genes including a potential prognostic biomarker IGSF3, highlighting NK-, B-cell mediated immunity, and B cell receptor signaling pathways missed by all spatial variable gene methods attempted. StarTrail, filling important gaps in current literature, enables deeper insights into tissue spatial architecture.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

S66: Best Student Paper Awards: Winners

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Symphony 1 (Lobby Level)

Organizer: Ran Tao, Vanderbilt University Medical Center
Chair: Siyuan Ma, Vanderbilt University Medical Center

5:00 PM–5:18 PM Speaker: Shuozhi Zuo, Colorado School of Public Health

Mediation analysis with the mediator and outcome missing not at random

Author(s): Shuozhi Zuo, Debashis Ghosh, Peng Ding, Fan Yang

Mediation analysis is widely used for investigating direct and indirect causal pathways through which an effect arises. However, many mediation analysis studies are challenged by missingness in the mediator and outcome. In general, when the mediator and outcome are missing not at random, the direct and indirect effects are not identifiable without further assumptions. We study the identifiability of the direct and indirect effects under some interpretable mechanisms that allow for missing not at random in the mediator and outcome. We evaluate the performance of statistical inference under those mechanisms through simulation studies and illustrate the proposed methods via the National Job Corps Study.

5:18 PM–5:18 PM Speaker: Yikun Zhang, University of Washington

Efficient inference on high-dimensional linear models with missing outcomes

Author(s): Yikun Zhang, University of Washington; Alexander Giessing, University of Washington; Yen-Chi Chen, University of Washington

High-dimensional data, where the number of covariates far exceeds the sample size, are pervasive in diverse domains, including genomics, quantitative finance, and healthcare studies. While the intricacies of high-dimensionality impose unusual challenges for conducting valid statistical inference, these challenges are further compounded when the outcome/response variable is potentially missing. In this talk, we propose an efficient debiasing method that addresses the above high-dimensional inference problem with missing outcomes. Specifically, we derive a debiased estimator by correcting the bias of a Lasso pilot estimate based on its weighted residuals. The weights are estimated by a convex debiasing program that trades off bias and variance optimally, which can be efficiently tuned and solved via its dual formulation. Provided that the propensity scores are consistently estimated by any machine learning methods, the proposed estimator is asymptotically normal and semi-parametrically efficient among all asymptotically linear estimators. We validate the finite-sample performance of our proposed estimator through comprehensive simulation studies and apply it to inferring the stellar masses of observed galaxies in the Sloan Digital Sky Survey. Finally, if time is allowed, we will briefly mention some potential applications of our proposed method to causal inference problems. The talk is based on a joint work with Alexander Giessing and Yen-Chi Chen.

5:36 PM–5:54 PM Speaker: Liying Chen, University of Michigan

GraphR: A probabilistic modeling framework for genomic networks incorporating sample heterogeneity

Author(s): Liying Chen, University of Michigan; Satwik Acharyya, University of Michigan; Chunyu Luo, University of Pennsylvania; Yang Ni, Texas A&M University; Veerabhadran Baladandayuthapani, University of Michigan

Probabilistic graphical models are powerful tools to infer, interpret, and visualize complex biological networks. However, most existing graphical models assume homogeneity across samples, limiting their application in heterogeneous contexts e.g. tumor and spatial heterogeneity. We propose a general and flexible Bayesian approach called Graphical Regression (GraphR) which incorporates intrinsic heterogeneity at different scales such as discrete, continuous and spatial, enables sparse network estimation at sample-specific level, has higher precision compared to existing approaches and is computationally efficient for analyses of large genomic datasets. We employ GraphR to analyze four diverse multi-omic and spatial transcriptomics datasets to infer inter- and intra-sample genomic networks and delineate several novel biological discoveries. We have also developed the GraphR R-package and a user-friendly Shiny App for analysis and dynamic network visualization.

5:54 PM–6:12 PM Speaker: Jiacheng Miao, University of Wisconsin-Madison

Assumption-lean and data-adaptive machine learning-assisted inference with application to genome-wide association studies

Author(s): Jiacheng Miao, University of Wisconsin-Madison; Yixuan Wu, University of Wisconsin-Madison; Zhongxuan Sun, University of Wisconsin-Madison; Xinran Miao, University of Wisconsin-Madison; Tianyuan Lu, University of Wisconsin-Madison; Jiwei Zhao, University of Wisconsin-Madison; Qiongshi Lu, University of Wisconsin-Madison

Machine learning (ML) has revolutionized analytical strategies in almost all scientific disciplines including human genetics and genomics. Due to challenges in sample collection and precise phenotyping, ML-assisted genome-wide association study (GWAS) which uses sophisticated ML to impute phenotypes and then performs GWAS on imputed outcomes has quickly gained popularity in complex trait genetics research. However, the validity of associations identified from ML-assisted GWAS has not been carefully evaluated. In this study, we report pervasive risks for false positive associations in ML-assisted GWAS. To address this issue, we introduce POP-Inf and POP-GWAS, two principled statistical frameworks for valid and powerful inference based on ML-predicted outcomes and ML-assisted GWAS. POP-Inf provides valid and powerful statistical inference irrespective of the quality of imputation or variables and algorithms used for ML imputation. It also only requires GWAS summary statistics as input. We employed POP-GWAS to perform the largest GWAS of bone mineral density (BMD) derived from dual-energy X-ray absorptiometry imaging at 14 skeletal sites, identifying 89 novel loci reaching genomewide significance and revealing skeletal site-specific genetic architecture of BMD. Our framework may fundamentally reshape the analytical strategies in future ML-assisted GWAS and ML-assisted inference. It also highlights the need for novel statistical methods alongside the rapid development of ML.

6:12 PM–6:30 PM Speaker: Chenyin Gao, North Carolina State University

Integrating randomized placebo-controlled trial data with external controls: A semiparametric approach with selective borrowing

Author(s): Chenyin Gao, North Carolina State University; Shu Yang, North Carolina State University; Mingyang Shan, Eli Lilly and Company; Wenyu Ye, Eli Lilly and Company; Ilya Lipkovich, Eli Lilly and Company; Douglas Faries, Eli Lilly and Company

In recent years, real-world external controls (ECs) have grown in popularity as a tool to empower randomized placebo-controlled trials (RPCTs), particularly in rare diseases or cases where balanced randomization is unethical or impractical. However, as ECs are not always comparable to the RPCTs, direct borrowing ECs without scrutiny may heavily bias the treatment effect estimator. Our paper proposes a data-adaptive integrative framework capable of preventing unknown biases of ECs. The adaptive nature is achieved by dynamically sorting out a set of comparable ECs via bias penalization. Our proposed method can simultaneously achieve (a) the semiparametric efficiency bound when the ECs are comparable and (b) selective borrowing that mitigates the impact of the existence of incomparable ECs. Furthermore, we establish statistical guarantees, including consistency, asymptotic distribution, and inference, providing type-I error control and good power. Extensive simulations and two real-data applications show that the proposed method leads to improved performance over the RPCT-only estimator across various bias-generating scenarios.

S67: CAUSAL INFERENCE AND SURVIVAL ANALYSIS

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Sound Emporium A/B (Mezzanine Level)

Organizer: Min Zhang, Tsinghua University
Chair: Di Wang, University of Michigan

5:00 PM–5:20 PM Speaker: Cheng Zheng, University of Nebraska Medical Center

Causal mediation analysis via joint modeling approach for multiple recurrent and terminal events

Author(s): Fang Niu, University of Nebraska Medical Center; Cheng Zheng, University of Nebraska Medical Center; Lei Liu, Washington University in St. Louis

Understanding the diverse causal mechanisms between primary exposure and outcomes has garnered significant interest in the social and medical fields. In the context of HIV patients, over 20 distinct opportunistic infections (OIs) present complex effects on the health trajectory and associated mortality. It is crucial to differentiate among these OIs to devise tailored strategies to enhance patients' survival and quality of life. However, existing statistical frameworks for studying causal mechanisms have limitations, either focusing on single mediators or lacking the ability to handle unmeasured confounding, especially for the survival outcomes. In this work, we propose a novel joint modeling approach that considers multiple recurrent events as mediators and survival endpoints as outcomes, relaxing the assumption of "sequential ignorability" by utilizing the shared random effect to handle unmeasured confounders. We assume the multiple mediators are not causally related to each other given observed covariates and the shared frailty. Simulation studies demonstrate good finite sample performance of our methods in estimating both model parameters and multiple mediation effects. We apply our approach to an AIDS study and evaluate the mediation effects of different types of OIs. We find that distinct pathways through the two treatments and CD4 counts impact overall survival via different types of recurrent opportunistic infections.

5:20 PM–5:20 PM Speaker: Yubai Yuan, Penn State University

De-confounding causal inference using latent multiple-mediator pathways

Author(s): Yubai Yuan, Annie Qu

Causal effect estimation from observational data is one of the essential problems in causal inference. However, most estimation methods rely on the strong assumption that all confounders are observed, which is impractical and untestable in the real world. We develop a mediation analysis framework inferring the latent confounder for debiasing both direct and indirect causal effects. Specifically, we introduce generalized structural equation modeling that incorporates structured latent factors to improve the goodness-of-fit of the model to observed data, and deconfound the mediators and outcome simultaneously. One major advantage of the proposed framework is that it utilizes the causal pathway structure from cause to outcome via multiple mediators to debias the causal effect without requiring external information on latent confounders. In addition, the proposed framework is flexible in terms of integrating powerful nonparametric prediction algorithms while retaining interpretable mediation effects. In theory, we establish the identification of both causal and mediation effects based on the proposed deconfounding method. Numerical experiments on both simulation settings and a normative

aging study indicate that the proposed approach reduces the estimation bias of both causal and mediation effects.

5:40 PM–6:00 PM Speaker: Liang Li, University of Texas MD Anderson Cancer Center

Propensity score analysis with local balance and calibration
Author(s): Maosen Peng, University of Texas MD Anderson Cancer Center; Yan Li, Mayo Clinic; Chong Wu, University of Texas MD Anderson Cancer Center; Liang Li, University of Texas MD Anderson Cancer Center

This presentation discusses a novel propensity score weighting analysis. We define two sufficient and necessary conditions for a function of the covariates to be the propensity score. The first is "local balance", which ensures the conditional independence of covariates and treatment assignment across a dense grid of propensity score values. The second condition, "local calibration," guarantees that a balancing score is a propensity score. Using three-layer feed-forward neural networks, we develop a nonparametric propensity score model that satisfies these conditions, effectively circumventing the issue of model misspecification and optimizing covariate balance to minimize bias and stabilize the inverse probability weights. Our proposed method performed substantially better than existing methods in extensive numerical studies of both real and simulated benchmark datasets.

6:00 PM–6:20 PM Speaker: Kevin (Zhi) He, University of Michigan

Incorporating external risk information with the time-to-event model under population heterogeneity
Author(s): Di Wang, University of Michigan; Kevin He, University of Michigan

Polygenic hazard scores (PHS) designed for European ancestry (EUR) individuals provide ample information regarding survival risk discrimination. Incorporating such information can improve the performance of risk discrimination in an internal small-sized non-EUR cohort. However, given that external EUR-based PHS and internal individual-level data come from different populations, ignoring population heterogeneity can introduce substantial bias. In this paper, we develop a Kullback-Leibler-based Cox model (CoxKL) to integrate internal individual-level time-to-event data with external risk scores derived from published prediction models. Partial-likelihood-based KL information is utilized to measure the discrepancy between the external risk information and the internal data. We establish the asymptotic properties of the CoxKL estimator. Simulation studies show that the integration model by the proposed CoxKL method achieves improved estimation efficiency and prediction accuracy. We apply the proposed method to develop trans-ancestry PHS models for prostate cancer and breast cancer and find that integrating EUR-based PHS with internal genotype data of African ancestry (AFR) individuals yields considerable improvement on the cancer risk discrimination.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

S68: INTEGRATIVE APPROACHES TO MULTI-OMICS DATA ANALYSIS FOR UNCOVERING DISEASE MECHANISMS

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Southern Ground A/B (Mezzanine Level)

Organizer: Qunhua Li, Penn State University
Chair: Xiang Zhu, Penn State University

5:00 PM–5:20 PM Speaker: Di Wu, University of North Carolina at Chapel Hill

Differential expression analysis by integrating metatranscriptomics and sample-paired metagenomics data

Author(s): Binghao Yan, University of Pennsylvania; Eunchong Kang, University of North Carolina at Chapel Hill; Di Wu, University of North Carolina at Chapel Hill

Microbial metatranscriptomics (MTX) is becoming increasingly important for profiling the gene expression pattern and functional activities of microbial communities. A fundamental task for analyzing the MTX data from high-throughput sequencing is the Differential Expression (DE) analysis, to identify up/down-regulated genes/pathways associated with the different sample groups (e.g., disease conditions). However, DE analysis is challenging in MTX due to the data distribution and how gene expression is affected by underlying DNA copies in metagenome (MGX) of the corresponding biological samples. Ignorance of the dependence between MTX and MGX in DE analysis may lead to false discoveries of differentially expressed genes. We proposed a new statistical framework, coNBMTX, in a conditional Negative Binomial regression model to identify genes in MTX only associated with sample groups but independent to MGX, by estimating sample group effects conditional on the information from sample-paired MGX data. We showed that the adjustment of underlying DNA copies could significantly eliminate the variations of RNA data, thus improving the power of statistical inference and giving meaningful biological interpretation. Using both the specifically designed simulated data and two microbial data sets, we demonstrated that our method has a high statistical power while controlling the false discovery rate. We applied our method to oral microbiome data and identified significant genes that are associated with Early Childhood Caries (ECC). Overall, our testing method enabled a more accurate and effective DE analysis based on MTX with sample-paired MGX data and helped us improve understanding of the functional characterization of microbial communities. Availability: https://github.com/Lizz647/coNBMTX.

5:20 PM–5:20 PM Speaker: Li-Xuan Qin, Memorial Sloan Kettering Cancer Center

Evidence-based practice for transcriptomic data harmonization

Author(s): Li-xuan Qin, Memorial Sloan Kettering Cancer Center

It is widely acknowledged that the reproducibility of transcriptomics data analysis depends on identifying and mitigating data artifacts that arise from disparate experimental handling through data harmonization. While data harmonization methods, encompassing data normalization and batch-effect correction, have proliferated to address these artifacts, there remains a scarcity of principled approaches to tailor the choice of a harmonization method to a specific dataset. Additionally, there is a paucity of statistical investigations placing data harmonization within the context of subsequent analyses. To advance evidence-based practices in omics data har-

monization, my research has produced robust benchmark data, novel statistical methods, and accompanying software tools, with a specific emphasis on microRNAs. In this presentation, I will introduce two recent methods we developed: (1) DANA (DAta-driven Normalization Assessment), a data-driven and biology-motivated approach for harmonization method selection, and (2) BatMan (BATch MitigAtion via stratificatioN), a batch-correction method specifically designed for survival analysis. We evaluated the performance of these methods using paired datasets from the same set of tumor samples collected at Memorial Sloan Kettering Cancer Center and applied them to additional datasets sourced from the Cancer Genome Atlas. R packages for these two methods can be found at https://github.com/LXQin.

5:40 PM–6:00 PM Speaker: Xuewei Cao, Columbia University

A gradient boosting informed multi-omics colocalization method improves the discovery of molecular quantitative trait loci for complex diseases

Author(s): Xuewei Cao, Columbia University; Haochen Sun, Columbia University; Ru Feng, Columbia University; Rahul Mazumder, Massachusetts Institute of Technology and the Alzheimer's Disease Functional Genomics Consortium; Kushal Dey, Memorial Sloan Kettering Cancer Center; Gao Wang, Columbia University

Background: In the emerging field of functional genomics research at population scale, the integration of multiple molecular quantitative trait loci (QTL) studies with GWAS is essential, especially with the widespread availability of multi-omics data like histone modification, gene expression, splicing, and protein abundance across various tissues and cell types. This need has highlighted the demand for advanced multi-trait colocalization methods. Current methods either lack sufficient accuracy in pinpointing causal variants or are computationally challenging for a wide range of molecular phenotypes. Method: To address this gap, we introduce ColocBoost, a method using a gradient boosting algorithm to identify shared putative causal variants in multi-omics QTL studies. We demonstrate that ColocBoost is powerful and calibrated compared to existing methods in real-world data-driven simulation studies involving multiple traits and variants in linkage disequilibrium. ColocBoost is capable of providing both variant and region level colocalization evidence. Results: When applied to integrate multi-omics QTL and Alzheimer's Disease (AD) GWAS datasets within the FunGen-xQTL Project, ColocBoost offers new insights into the functional roles of xQTL in AD etiology. This method not only offers an adaptable framework for functional signal detection in genetic research but also advances multi-trait colocalization techniques, effectively refining disease GWAS findings with vast molecular QTL data resources.

6:00 PM–6:20 PM Speaker: Xiang Zhu, Penn State University

Localizing disease-causing genes through the lens of enhancers

Author(s): Xiang Zhu, Penn State University

Over the past two decades, genome-wide association studies (GWAS) have advanced the discovery and interpretation of numerous genetic variants influencing complex human diseases. Notably, over 90% of disease-associated variants reside in non-coding regions of the human genome, which complicates the identification of causal genes underpinning most GWAS discoveries. Enhancers, a major group of non-coding elements that interact with transcription factors to increase gene expression levels, can offer crucial insights into the regulatory mechanisms

by which non-coding mutations influence downstream genes and disease phenotypes. Building on this premise, we present an enhancer-centric approach to improve the discovery and interpretation of effector genes for complex diseases in GWAS. First, we develop a Bayesian hierarchical modeling framework to prioritize the genetic associations of enhancer-like sequences with a specific disease, based on the corresponding GWAS data. Next, we identify sequence motifs that are enriched in a target set of disease-associated enhancers compared to a background set of non-associated enhancers, while accounting for the sequence and functional features of both sets. Lastly, we shortlist disease-associated enhancers that contain the most enriched motifs, and then link these enhancers to their putative target genes through a whole host of functional resources such as 3D genome profiling and CRISPR screening. Extensive simulations confirm that our approach is well-calibrated. Applying this approach to real-world GWAS highlights multiple candidate effector genes that are functionally relevant to obesity and schizophrenia but were not identified by the same GWAS, nor by updated GWAS with much larger sample sizes, nor by existing multi-omics analysis of these GWAS data. In summary, this work underscores the potential of enhancers as a useful tool to prioritize causal genes underpinning GWAS findings, and showcases a generalizable strategy to realize this potential.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

S69: ADDRESSING HETEROGENEITY IN FEDERATED AND TRANSFER LEARNING IN REAL-WORLD DATA

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Blackbird B (Mezzanine Level)

Organizer: Yong Chen, University of Pennsylvania Chair: Yudong Wang, University of Pennsylvania

5:00 PM–5:20 PM Speaker: Chongliang Luo, Washington University in St. Louis

One-shot distributed algorithm for multi-center Cox model with time-varying coefficients

Author(s): C. Jason Liang, National Institute of Allergy and Infectious Diseases; Chongliang Luo, Washington University School of Medicine; Henry Kranzler, University of Pennsylvania; Jiang Bian, University of Florida; Yong Chen, University of Pennsylvania

The Cox model with time-varying coefficients relaxes the proportional hazards assumption of the usual Cox model but requires additional data to accurately estimate time-varying coefficients. Electronic health records (EHR) often contain sufficient data for the relaxed Cox model to be feasible. However, individual health systems have limited sample sizes and large-scale EHR data networks are proliferating. Nevertheless, the data in these networks are decentralized as sharing of individual-level data is often unpractical due to patient privacy concerns, and therefore require the use of a distributed algorithm to utilize all the available data. We propose a One-shot Distributed Algorithm to fit multi-center Cox regression models with Time-varying coefficients (ODACT). Our method precisely estimates the time-varying effects over time, leading to statistical inference that is nearly as precise as analysis by direct pooling of the decentralized EHR data. We applied our method to study risk factors associated with opioid use disorder (OUD) after chronic non-cancer pain (CNCP) opioid prescription using decentralized data from a large clinical research network across 5 sites with 69,163 subjects.

5:20 PM–5:20 PM Speaker: Xiaokang Liu, University of Missouri

Targeted learning via probabilistic subpopulation matching

Author(s): Xiaokang Liu, University of Missouri; Jie Hu, University of Pennsylvania; Naimin Jing, Merck & Co., Inc.; Runze Li, Pennsylvania State University; Yong Chen, University of Pennsylvania

In biomedical research, to obtain more accurate prediction results from a target study, leveraging information from multiple similar source studies is proved to be useful. However, in most biomedical applications based on real-world data, populations under consideration in different studies (e.g., clinical sites) can usually be heterogeneous, leading to significant differences across studies. Traditional methods are typically based on study matching to identify source studies similar to the target study, and samples from source studies that significantly differ from the target study will all be dropped, which can lead to potential loss of information. We consider a situation where each source study has a large population containing subgroups of people that are similar to the target study, which gives rise to the idea of target learning via subpopulation matching instead of study matching. Measuring similarities between subpopulations can effectively decompose potentially large between-study heterogeneity and therefore allows incorporating information from all source studies to aid the target learning. We devise the method as a two-step procedure, where a finite mixture model is first fitted jointly across all studies to provide subject-wise probabilistic subpopulation information, followed by a step of within-subpopulation information transferring from source studies to the target study for

each identified subpopulation. We establish the non-asymptotic properties of our estimator and demonstrate the ability of our method to improve prediction at the target study via simulation studies. We apply our method to predict the occurrence of long-COVID for children at a children's hospital by leveraging information from other children's hospitals.

5:40 PM–6:00 PM Speaker: Qiong Wu, University of Pennsylvania

Bias reduction in comparative effectiveness research of COVID-19 vaccines
Author(s): Qiong Wu, University of Pennsylvania; Huiyuan Wang, University of Pennsylvania; Jiayi Tong, University of Pennsylvania; Christopher Forrest, Children's Hospital of Philadelphia; Jeffrey Morris, University of Pennsylvania; Yong Chen, University of Pennsylvania

Post-introduction evaluation of COVID-19 vaccines is essential to answer critical questions not fully addressed in clinical trials, such as effectiveness against evolving SARS-CoV-2 variants and the incidence of rare adverse events, and thus inform vaccine recommendations to the general public. However, in the U.S., the fragmentation of immunization records across multiple, often disconnected, systems results in incomplete vaccination histories in patients' Electronic Health Records (EHRs). Such incomplete treatment information can lead to biased estimates in the comparative effectiveness research of COVID-19 vaccines. To address this, we consider a setting where internal validation is achieved with reliable vaccination records by integrating EHR data with immunization information systems. We introduce the identifying formula and the efficient influence function for average treatment effects in contexts where such internal validation is available. Following this, we construct an efficient estimator that achieves full statistical efficiency and has faster rates of convergence than estimating complex nuisance parameters. The application of the proposed method on data from a national network of U.S. pediatric medical centers (PEDSnet) assessed the effectiveness of vaccination among children and adolescents during the Delta and Omicron periods.

6:00 PM–6:20 PM Speaker: Huiyuan Wang, University of Pennsylvania

Robust and efficient high-dimensional inference with surrogate outcomes
Author(s): Huiyuan Wang, University of Pennsylvania; Yang Ning, Cornell University; Yong Chen, University of Pennsylvania

Electronic health records (EHR) offer a valuable resource for discovering novel disease risk factors. However, the common issue of missingness in the primary phenotype of interest often leads to efficiency loss in inferential methods that rely solely on fully observed samples. Additionally, the prevalent misclassification of EHR-derived phenotypes can result in systematic bias, thereby affecting the reproducibility of findings. In response to these challenges, our study introduces a robust and efficient framework for high-dimensional EHR-based discovery. Based on a class of surrogate models for EHR-based phenotypes, we construct an augmented score function and develop a corresponding test statistic. The statistic not only maintains correct coverage under the null hypothesis but also exhibits enhanced power under local alternatives, outperforming tests that only use fully observed samples. Surprisingly, it achieves the correct coverage even in scenarios with arbitrary misclassification of EHR-based phenotypes and misspecified surrogate models. The statistical effectiveness of our proposed method is evaluated through extensive simulations and a real-world data application.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

S70: RECENT ADVANCES IN STATISTICAL ANALYSIS OF NETWORK AND MATRIX DATA: SPECTRAL METHODS AND BEYOND

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Gold (Lower Level)

Organizer: Fangzheng Xie, Indiana University
Chair: Runbing Zheng, Johns Hopkins University

5:00 PM–5:20 PM Speaker: Shujie Ma, University of California, Riverside

Privacy-preserving community detection for locally distributed multiple networks

Author(s): Shujie Ma, University of California, Riverside

In this talk, I will introduce a new efficient and scalable consensus community detection approach and distributed learning algorithm in a multi-layer stochastic block model using locally stored network data with privacy-preserving. Specifically, we develop a spectral clustering-based algorithm named ppDSC. To reduce the bias incurred by the randomized response (RR) mechanism for achieving differential privacy, we develop a two-step bias adjustment procedure. To reduce the communication cost encountered in distributed learning, we perform the eigen-decomposition locally and then aggregate the local eigenvectors using an orthogonal Procrustes transformation. We establish a novel bound on the misclassification rate of ppDSC. The new bound reveals the asymmetric roles of the two edge-flipping probabilities of the RR in the misclassification rate. Through the bound, we can also find the optimal choices for the flipping probabilities given a fixed privacy budget. Moreover, we show that ppDSC enjoys the same statistical error rate as its centralized counterpart, when the number of machines satisfies a polynomial order with the sample size on each local machine and the effective heterogeneity is well controlled.

5:20 PM–5:20 PM Speaker: Xiaodong Li, University of California, Davis

Rank selection for weighted degree-corrected stochastic block models

Author(s): Xiaodong Li, University of California, Davis

We will present our recent work on how to select the number of communities for weighted networks, particularly for count-weights. We first extend the degree-corrected stochastic block model (DCSBM) to weighted networks in order to model the mean structure. Moreover, we assume the relationship between the entrywise variances and corresponding means can be characterized a variance function. Our procedure follows a sequential testing framework. For each candidate number of communities, denoted as m, we carry our spectral clustering and then estimate the mean structure. The variance profile matrix is then estimated through the variance function. We then normalize the adjacency matrix through carrying out Sinkorn scaling of the estimated variance profile matrix, and a spectral statistic based on the normalized adjacency matrix is proposed to determine whether m is the true number of communities. We establish consistency result for our procedure for sparse networks, where the analysis relies on some recent progress in random matrix theory. Extensive numerical experiments on simulated and real network data also illustrate the competitive empirical properties of our procedure.

5:40 PM–6:00 PM Speaker: Fangzheng Xie, Indiana University

Higher-order entrywise eigenvectors analysis of low-rank random matrices: Bias correction, Edgeworth expansion, and bootstrap

Author(s): Fangzheng Xie, Indiana University; Yichi Zhang, Duke University

Understanding the distributions of spectral estimators in low-rank random matrix models, also known as signal-plus-noise matrix models, is fundamentally important in various statistical learning problems, including network analysis, matrix denoising, and matrix completion. In this talk, we will investigate the entrywise eigenvector distributions in a broad range of low-rank signal-plus-noise matrix models by establishing their higher-order accurate stochastic expansions. At a high level, the stochastic expansion states that the eigenvector perturbation approximately decomposes into the sum of a first-order term and a second-order term, where the first-order term in the expansion is a linear function of the noise matrix, and the second-order term is a linear function of the squared noise matrix. Our theoretical finding is used to derive the bias correction procedure for the eigenvectors. We further establish the Edgeworth expansion formula for the studentized entrywise eigenvector statistics. In particular, under mild conditions, we show that Cramér's condition on the smoothness of noise distribution is not required, thanks to the self-smoothing effect of the second-order term in the eigenvector stochastic expansion. The Edgeworth expansion result is then applied to justify the higher-order correctness of the residual bootstrap procedure for approximating the distributions of the studentized entrywise eigenvector statistics.

6:00 PM–6:20 PM Speaker: Jingming Wang, Harvard University

Optimal network membership estimation under severe degree heterogeneity

Author(s): Zheng Tracy Ke, Harvard University; Jingming Wang, Harvard University

Real networks often have severe degree heterogeneity, with maximum, average, and minimum node degrees differing significantly. In this talk, I will discuss the impact of degree heterogeneity on statistical limits of network data analysis. Introducing the empirical heterogeneity distribution (EHD) under a degree-corrected mixed membership model, we show that the optimal rate of mixed membership estimation is an explicit functional of the EHD. This result confirms that severe degree heterogeneity decelerates the error rate, even when the overall sparsity remains unchanged. To obtain a rate-optimal method, we modify an existing spectral algorithm, Mixed- SCORE, by adding a pre-PCA normalization step. This step normalizes the adjacency matrix by a diagonal matrix consisting of the b-th power of node degrees. We discover that b = 1/2 is universally favorable. The resulting spectral algorithm is rate-optimal for networks with arbitrary degree heterogeneity. A technical component in our proofs is entry-wise eigenvector analysis of the normalized graph Laplacian. This is a joint work with Zheng Tracy Ke.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

S71: RECENT ADVANCES IN COMPLEX SURVIVAL DATA ANALYSIS WITH NOVEL APPLICATIONS

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Melody (Lobby Level)

Organizer: Yichuan Zhao, Georgia State University
Chair: Dongmei Li, University of Rochester Medical Center

5:00 PM–5:20 PM Speaker: Zhigang Li, University of Florida

Joint modeling in presence of informative censoring on the retrospective time scale with application to palliative care research

Author(s): Quran Wu, University of Florida; Michael Daniels, University of Florida; Zhigang Li, University of Florida

Joint modeling of longitudinal data such as quality of life data and survival data is important for palliative care researchers to draw efficient inferences because it can account for the associations between those two types of data. Modeling quality of life on a retrospective from death time scale is useful for investigators to interpret the analysis results of palliative care studies which have relatively short life expectancies. However, informative censoring remains a complex challenge for modeling quality of life on the retrospective time scale although it has been addressed for joint models on the prospective time scale. To fill this gap, we develop a novel joint modeling approach that can address the challenge by allowing informative censoring events to be dependent on patients' quality of life and survival through a random effect. There are two sub-models in our approach: a linear mixed effect model for the longitudinal quality of life and a competing-risk model for the death time and dropout time that share the same random effect as the longitudinal model. Our approach can provide unbiased estimates for parameters of interest by appropriately modeling the informative censoring time. Model performance is assessed with a simulation study and compared with existing approaches. A real-world study is presented to illustrate the application of the new approach.

5:20 PM–5:20 PM Speaker: Jing Xu, University of North Carolina at Charlotte

Generalized semiparametric intensity models for recurrent event data with applications

Author(s): Jing Xu, University of North Carolina at Charlotte; Yanqing Sun, University of North Carolina at Charlotte; Fei Heng, University of North Florida; Peter B. Gilbert, University of Washington

This research is motivated by the MAL094 malaria vaccine efficacy trial aimed to test the efficacy of the RTS,S/AS01E malaria vaccine. We studied a class of generalized semiparametric intensity models for recurrent events. The models feature unspecific time- varying effects and constant effects, while the effects that depend on time-varying covariates or event history are modeled parametrically. The models offer flexibility through the selection of different link functions and parametric functions. Estimation involves local linear estimation and profile log-likelihood estimation. The asymptotic properties of estimators are developed using martingale theory and empirical processes. Two hypothesis tests based on residual processes are developed to assess the parametric functions of the covariate-varying effects. Simulation studies show that the methods perform well in finite samples. The proposed methods are applied to MAL094 malaria vaccine efficacy trial data to investigate how malaria infection risk depends on previous

infections, vaccinations as well as other factors.

5:40 PM–6:00 PM Speaker: Song Yang, National Heart, Lung, and Blood Institute (NHLBI), NIH

Semi- and non-parametric inference of treatment effects with terminal and non-terminal events when both are subject to competing risks

Author(s): Song Yang, National Heart, Lung, and Blood Institute (NHLBI), NIH

In clinical trials with time to event outcomes, often the outcomes of interest are a terminal event such as the cardiovascular death, and one or more non-terminal events such as heart failure hospitalization. Traditionally time-to-first-event analysis is used with the log-rank test or hazard ratio estimator derived under the Cox model for the first event. However, this approach uses data inefficiently. In recent years, more statistical models and procedures have been developed using effect measures such as the restricted mean survival time and its variants, the win ratio and its variants, and the Wei-Lachin test, as well as multivariate models such as the copula models and the illness-death models. In this talk, a few new semi- and non-parametric Inference procedures are proposed. The data structure and testing and estimation problems are formulated to accommodate competing risks for both the terminal event and the non-terminal events. The presence of competing risks can otherwise lead to biased results and obscure interpretations. The new methods can be rigorously justified and they are illustrated using data from a few recent large cardiovascular trials.

6:00 PM–6:20 PM Speaker: Ali Jinnah, Georgia State University

Jackknife empirical likelihood methods for the Cox regression model

Author(s): Ali Jinnah, Georgia State University

In this paper, we propose jackknife empirical likelihood methods to draw inference for the regression parameters in Cox regression model. We develop the jackknife empirical likelihood (JEL), adjusted jackknife empirical likelihood (AJEL), mean jackknife empirical likelihood (MJEL), transformed jackknife empirical likelihood (TJEL), and adjusted transformed jackknife empirical likelihood (TAJEL) methods. We profile the set of nuisance parameters to study the parameter of interest. Extensive simulation studies show that the proposed jackknife empirical likelihood methods have better performance than the normal approximation in most cases particularly for data with high censoring rates. We apply the proposed methods to study the Bone Marrow Transplant Patients (BMT), Larynx, and Myeloma real datasets for illustration.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

S72: RECENT ADVANCES IN ANALYSIS: NATURAL LANGUAGE PROCESSING OF MENTAL HEALTH AND METHODS FOR MICROBIOME DATA

Tuesday, June 18, 2024
5:00 PM–6:30 PM, Green Room (Lobby Level)

Organizer: Yimei Li, University of Pennsylvania Chair: Jing Ma, Fred Hutchinson Cancer Center

5:00 PM–5:20 PM Speaker: Jing Ma, Fred Hutchinson Cancer Center

Structured dimensionality reduction of multi-view data

Author(s): Jing Ma, Fred Hutchinson Cancer Center

Exploratory analysis of multiple datasets measured on the same set of samples can generate insights that are not immediately obvious when analyzing a single dataset. Canonical correlation analysis (CCA) is commonly applied, but CCA and its variants ignore structures present in the datasets and may not provide the most interpretable results. To address these limitations, we introduce a new framework of integrative principal component analysis that allows the analyst to incorporate side information about the relationships among samples and relationships among variables. This leads to a shared low-dimensional representation of the samples which both describes the latent structure and has interpretable axes. Furthermore, the method can uncover latent structures unique to each dataset. We show that our method does well at reconstructing the latent structure in simulated data and also demonstrate its performance on a real data application.

5:20 PM–5:20 PM Speaker: Jiang Gui, Dartmouth College

Words matter: An association study between natural language processing of clinical mental health notes and suicide risk.

Author(s): Jiang Gui, Dartmouth College; Siting Li, Dartmouth College; Max Levis, Veterans Affairs Medical Center, White River Junction, VT (VA White River Junction); Monica DiMambro, VA White River Junction; Weiyi Wu, Dartmouth College; Joshua Levy, Cedars Sinai Medical Center; Brian Shiner, VA White River Junction

Introduction: Suicide risk assessment relies heavily on clinical evaluation and patient self-report, which have limitations. Natural language processing (NLP) of electronic health records provides an alternative approach to extracting informative risk predictors from clinical notes. However, modeling NLP variables is challenged by zero inflation and skewed distributions.

Materials and Methods: This study evaluated whether an adaptive mixture categorization (AMC) method could optimize the suicide risk predictive capacity of NLP data extracted from Veterans Health Administration clinical notes. NLP variables for 25,637 patients were generated using the SÉANCE software. AMC identified data-driven thresholds to categorize NLP measures into distinct groups maximizing between-category variance. Associations between suicide outcome and AMC-transformed NLP predictors were compared to original and quantile-categorized NLP variables. Results: AMC-processed variables showed stronger associations with suicide risk versus other approaches in full cohort analysis and sensitivity analyses using under-sampled data. Additionally, over 90% of the NLP variables are significantly associated with suicide risk.

Conclusions: AMC-based categorization substantially enhanced the suicide predictive capacity of NLP variables extracted from clinical text. Transforming skewed NLP data with AMC holds promise for improving risk prediction models integrating natural language information from electronic medical records.

5:40 PM–6:00 PM Speaker: Jun Chen, Mayo Clinic

mPower: a real data-based power analysis tool for microbiome study design
Author(s): Jun Chen, Mayo Clinic

Power analysis is a critical step in designing a microbiome study. Previous power calculation tools mainly rely on parametric models, which underestimate the complexity of microbiome data and could produce overly optimistic power estimates. In this work, we present a simulation-based power analysis tool, mPower, to facilitate realistic power calculation for microbiome studies. The tool uses a real data-based semi-parametric simulation framework to generate realistic microbiome data, upon which the power assessment is performed. Coupled with our recently developed differential analysis tool, ZicoSeq, our power tool supports different study designs, including cross-sectional, case-control, and longitudinal studies, with or without confounders. It allows power analysis for both community-level and taxa-level testing. By using a database of large microbiome datasets from different environments, the users could perform power calculations based on the environment of interest. We will showcase the application of our power analysis tool using several real examples.

6:00 PM–6:20 PM Speaker: Pixu Shi, Duke University

Dimension reduction of longitudinal microbiome data
Author(s): Pixu Shi, Duke University

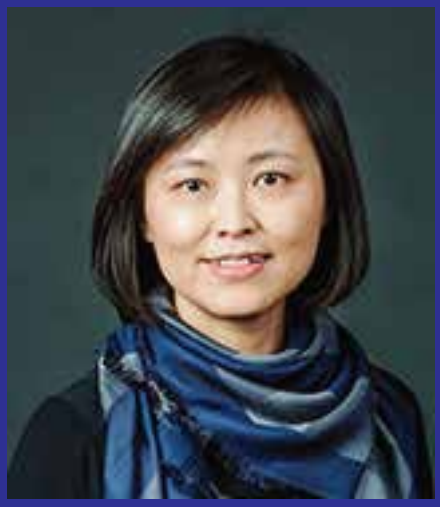
Complex dynamics of microbial communities underlie their essential roles in health and disease, but our understanding of these dynamics remains incomplete. To bridge this gap, longitudinal microbiome data are being rapidly generated, yet their power is limited by technical challenges in design and analysis, such as varying temporal sampling, noisy temporal patterns, complex correlation structures over feature and time, and high dimensionality. In this talk, we will first present TEMPoral TEnsor Decomposition (TEMPTED), the only time-informed dimensionality reduction method that extracts the underlying microbial dynamics while overcoming the statistical challenges posed by this type of data. TEMPTED extracts major temporal dynamics and key contributing features, facilitates beta-diversity analysis at both sample and subject levels, and promotes reproducibility by enabling the transfer of the learned low-dimensional representation from training data to unseen test data.

6:20 PM–6:30 PM **Q&A and Floor Discussion**

Banquet and Awards Ceremony – Wednesday, June 19

Symphony 2&3 (Lobby Level)
Ticket required

7:00–7:40 pm	Banquet Check-In and Buffet Dinner
7:40–8:40 pm	Banquet Speech by Dr. Mingyao Li, FIMS, FASA, FAAAS



From statistical genetics and genomics to precision pathology: Navigating the shift with machine learning and AI

Dr. Mingyao Li received her PhD in Biostatistics from the University of Michigan in 2005. She was trained as a statistical geneticist, but since she joined the faculty at the University of Pennsylvania in 2006, she has gradually transitioned her research from traditional statistical genetics to statistical genomics to have a deeper understanding of the molecular mechanism of human disease.

The central theme of her current research is to use statistical and machine learning methods to understand cellular heterogeneity in human-disease-relevant tissues, to characterize gene expression diversity across cell types, to study the patterns of cell state transition and crosstalk of various cells using data generated from single-cell and spatial transcriptomics studies, and to translate these findings into clinics. More recently, she expanded her expertise into computational pathology, which is critical when processing and analyzing spatial transcriptomics data. In addition to methods development, she is also interested in collaborating with researchers seeking to identify complex disease susceptibility genes and acting cell types. She has published extensively in cardiovascular disease and age-related macular degeneration, and more recently in cancer. At the University of Pennsylvania, she serves as the Director of the Statistical Center for Single-Cell and Spatial Genomics.

8:40–8:50 pm	Best SIBS Paper Awards presented by Dr. Hongkai Ji, Editor in Chief, *Statistics in Biosciences*
8:50–9:10 pm	Best Student Paper Awards presented by Dr. Siyuan Ma, co-chair, Student Paper Competition Committee
9:10–9:20 pm	Best Student Poster Awards presented by Dr. Fei Ye, chair, Poster Session Committee

Wednesday, June 19, 2024

7:30 – 11:00 am	Registration	Symphony Foyer Entrance (Lobby Level)
7:30 – 8:30 am	Breakfast	Symphony Foyer (Lobby Level)
8:30 – 9:30 am	Keynote: Jing Huang, PhD, FASA	Symphony 2&3 (Lobby Level)
9:30 – 10:00 am	Coffee Break	Symphony Foyer (Lobby Level)
10:00 – 11:30 am	Invited Sessions	

Online agenda: symposium2024.icsa.org/detailed-agenda

Ideas for lunch: symposium2024.icsa.org/nashville

Share your experience on social media! #ICSAnashville

Keynote Talk – Wednesday, June 19

CEO of your career: Mastering the art of charting your own fulfilling path in statistics

Jing Huang, PhD, FASA

8:30 am, Symphony 2&3 (Lobby Level)

Join us for an engaging exploration that is tailored to help each attendee carve out a fulfilling career in the vibrant field of statistics. Our journey begins with "Demand"—we identify the pressing societal needs that statisticians are uniquely equipped to meet, emphasizing how you can align your career with your inner drive to make a meaningful impact. Next, we pivot to "Expertise," where we discuss the crucial skills that extend beyond traditional mathematics, statistics, and data science—skills that are vital for advancing a rewarding and impactful career. Finally, we explore "Passion." What truly motivates you and helps you overcome challenges? How do you pinpoint your passions when data and experience are scarce? By synthesizing these three elements—demand, expertise, and passion—this presentation delivers a personalized blueprint for achieving both individual fulfillment and professional success in the dynamic world of statistics.

Dr. Jing Huang is Senior Vice President of Bioinformatics & Data Science at Veracyte, a molecular diagnostic company. Her division is responsible for creating, implementing, and executing bioinformatics pipelines, algorithm development, and statistical analyses across all phases of product development. Dr. Huang received her BA in Statistics and Probability from Peking University and her PhD in Statistics and MS in Epidemiology from Stanford University. She has been working in the biomedical field for over 20 years, and her research interest focuses on statistical methodologies in clinical trial design, genomic analysis, and machine learning. Dr. Huang has co-authored more than 30 articles in peer-reviewed scientific journals with nearly ten thousand citations and is a co-inventor of over 20 patent filings. Dr. Huang was elected as an ASA Fellow in 2023 to recognize her outstanding contributions to the medical research community in the field of statistics, for numerous statistical innovations in genomic tests, and for exemplary leadership and community service to the profession. Besides her daily work, she actively promotes data science through many of her volunteer activities. These include serving as is the founding president of DahShu, a 501(c)(3) nonprofit organization with the mission of promoting research and education in data science. She is currently the chapter representative of the ASA's San Francisco Bay Area Chapter (SFASA) and has served the organization for many years in various roles, including president and vice president of biostatistics. She was the general co-chair and local organization chair of the Fourteenth Asia Pacific Bioinformatics Conference (APBC 2016) as well as the DahShu 2017 Scientific Symposium on Computational Precision Health (CPH 2017).

S73: STATISTICAL METHODS FOR DATA FUSION AND THEIR APPLICATIONS

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Blackbird A (Mezzanine Level)

Organizer: Rui Duan, Harvard University
Chair: Rui Duan, Harvard University

10:00 AM–10:20 AM Speaker: Alex Luedtke, University of Washington

Probabilistic programming of efficient estimators

Author(s): Alex Luedtke, University of Washington

We introduce an algorithm that simplifies the construction of efficient estimators, making them accessible to a broader audience. "Dimple" takes as input computer code representing a parameter of interest and outputs an efficient estimator. Unlike standard approaches, it does not require users to derive a functional derivative known as the efficient influence function. Dimple avoids this task by applying automatic differentiation to the statistical functional of interest. Doing so requires expressing this functional as a composition of primitives satisfying a novel differentiability condition. Dimple also uses this composition to determine the nuisances it must estimate. In software, primitives can be implemented independently of one another and reused across different parameters of interest. We provide a preliminary Python implementation and showcase through examples how it allows users to go from parameter specification to efficient estimation with just a few lines of code. Applications to data fusion problems are discussed.

10:20 AM–10:20 AM Speaker: Zehang Li, University of California, Santa Cruz

Domain adaptive cause-of-death assignment using verbal autopsies under distribution shift

Author(s): Zehang Li, University of California, Santa Cruz

Understanding cause-specific mortality rates is crucial for monitoring population health and designing public health interventions. Worldwide, two-thirds of deaths do not have a cause assigned. Verbal autopsy (VA) is a well-established tool to collect information describing deaths outside of hospitals by conducting surveys to caregivers of a deceased person. It is routinely implemented in many low- and middle-income countries. Statistical algorithms to assign cause of death using VAs are typically vulnerable to the distribution shift between the data used to train the model and the target population. This presents a major challenge for analyzing VAs as labeled data are usually unavailable in the target population. This talk discusses a latent class model framework for VA data that jointly models VAs collected over heterogeneous domains, such as multiple study sites, different time periods, or distinct subpopulation. We introduce a parsimonious representation of the joint distribution of the collected symptoms and develop a computationally efficient algorithm for posterior inference and out-of-domain cause-of-death assignment. We will also discuss the importance of accounting for data shift in other related decision-making problems in VA studies.

10:40 AM–11:00 AM Speaker: Fatema Shafie Khorassani, Boston University

Data fusion for studying cancer surveillance databases

Author(s): Fatema Shafie Khorassani, Boston University; Jeremy Taylor, University of Michigan; Xu Shi, University of Michigan

Studying racial disparities in cancer mortality requires data on variables including healthcare access, socioeconomic status, and comorbidities. Existing national cancer surveillance databases each collect parts of the needed information. Data fusion allows us to study associations between race and cancer mortality adjusted for important confounders. Our goal is to make inference about a model regressing on covariates that come from two separate sources. The outcome of interest is collected in one dataset, and a set of important confounders are collected in another, and both sources have some common variables. We propose a data fusion method for time-to-event outcomes and derive semiparametric estimating equations for data fusion. Our proposed estimating equation is doubly robust, providing consistent parameter estimates if either the source process or the distribution of unobserved covariates is correctly specified. We apply the estimating equations to studying racial disparities in cancer mortality using data from the National Cancer Institute's Surveillance, Epidemiology, and End Results registry adjusted for confounders collected in the National Cancer Database.

11:00 AM–11:20 AM Speaker: Boyu Ren, McLean Hospital

Cross-study learning for generalist and specialist predictions

Author(s): Boyu Ren, McLean Hospital; Prasad Patil, Boston University; Francesca Dominici, Harvard University; Christian Webb, McLean Hospital; Giovanni Parmigiani, Harvard University; Lorenzo Trippa, Harvard University

The integration and use of data from multiple studies, for the development of prediction models is an important task in several scientific fields. We propose a framework for generalist and specialist predictions that leverages multiple datasets, with potential differences in the relationships between predictors and outcomes. Our framework uses stacking, and it includes three major components: (1) an ensemble of prediction models trained on one or more datasets, (2) task-specific utility functions and (3) a no-data-reuse technique for estimating stacking weights. We illustrate that under mild regularity conditions the framework produces stacked PFs with oracle properties. In particular we show that the stacking weights are nearly optimal. We also characterize the scenario where the proposed no-data-reuse technique increases prediction accuracy compared to stacking with data reuse in a special case. We perform a simulation study to illustrate these results and apply our framework to predict (1) mortality using a collection of datasets on long-term exposure to air pollutants and (2) elevated negative emotions using passive smartphone sensor data.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S74: RECENT ADVANCES IN PRECISION MEDICINE AND ADAPTIVE EXPERIMENTS

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Lyric (Lobby Level)

Organizer: Donglin Zeng, University of Michigan
Chair: Peijun Sang, University of Waterloo

10:00 AM–10:20 AM Speaker: Jingshen Wang, University of California, Berkeley

Adaptive experiments toward learning treatment effect heterogeneity

Author(s): Waverly Wei, University of Southern California; Xinwei Ma, UCSD; Jingshen Wang, UC Berkeley

Understanding treatment effect heterogeneity has become an increasingly popular task in various fields, as it helps design personalized advertisements in e-commerce or targeted treatment in biomedical studies. However, most of the existing work in this research area focused on either analyzing observational data based on strong causal assumptions or conducting post hoc analyses of randomized controlled trial data, and there has been limited effort dedicated to the design of randomized experiments specifically for uncovering treatment effect heterogeneity. In the manuscript, we develop a framework for designing and analyzing response adaptive experiments toward better learning treatment effect heterogeneity. Concretely, we provide response adaptive experimental design frameworks that sequentially revise the data collection mechanism according to the accrued evidence during the experiment. Such design strategies allow for the identification of subgroups with the largest treatment effects with enhanced statistical efficiency. The proposed frameworks not only unify adaptive enrichment designs and response-adaptive randomization designs but also complement A/B test designs in e-commerce and randomized trial designs in clinical settings. We demonstrate the merit of our design with theoretical justifications and in simulation studies with synthetic e-commerce and clinical trial data.

10:20 AM–10:20 AM Speaker: Young-geun Kim, Columbia University

Deep identifiable generative models for multi-modal data analysis

Author(s): Young-geun Kim, Columbia University; Ying Liu, Columbia University

The advent of large-scale multi-site data has brought opportunities to understand individuals' disease progression better. However, their high-dimensional and multi-modal nature presents significant challenges. We present a deep learning-based identifiable representation learning method for multi-modal biomedical data to fill this gap. Our approach builds upon the foundation of identifiable variational autoencoders (iVAEs), which allow the introduction of clinical covariates into nonlinear dimension reduction while deep generative models identify representations. Representations from iVAEs are interpretable with clinical covariates, but they are often entangled in the multi-modal data analysis, hindering the interpretation of their roles for each modality. To address this limitation, we extend iVAEs by decomposing representations into shared and individual components based on modality dependencies. Distinct from existing approaches, our proposed method uses nonlinear mapping to extract disentangled representations reflecting multi-modal structures while representations are clustered well by clinical covariates.

Experiments on multi-site data demonstrate the efficacy of the proposed method.

10:40 AM–11:00 AM Speaker: Qi Xu, University of California, Irvine

Multi-label residual weighted learning for individualized combination treatment rule
Author(s): Qi Xu, University of California Irvine; Xiaoke Cao, University of California Irvine; Geping Chen, Iowa State University; Hanqi Zeng, Harvard University; Haoda Fu, Eli Lilly and Company; Annie Qu, University of California Irvine

Individualized treatment rules (ITRs) have been widely applied in many fields such as precision medicine and personalized marketing. Beyond the extensive studies on ITR for binary or multiple treatments, there is considerable interest in applying combination treatments. This paper introduces a novel ITR estimation method for combination treatments incorporating interaction effects among treatments. Specifically, we propose the generalized ψ-loss as a non-convex surrogate in the residual weighted learning framework, offering desirable statistical and computational properties. Statistically, the minimizer of the proposed surrogate loss is Fisher-consistent with the optimal decision rules, incorporating interaction effects at any intensity level—a significant improvement over existing methods. Computationally, the proposed method applies the difference-of-convex algorithm for efficient computation. Through simulation studies, we demonstrate the superior performance of the proposed method in recommending combination treatments.

11:00 AM–11:20 AM Speaker: Daiqi Gao, Harvard University

Sequential decision making with sparsely observed rewards
Author(s): Daiqi Gao, Harvard University; Hsin-Yu Lai, Harvard University; Susan Murphy, Harvard University

Cardiac rehabilitation (CR) is an outpatient risk-reduction program for patients with cardiovascular disease, involving supervised exercise, dietary counseling, and stress management. However, a significant challenge remains in sustaining the heart-healthy lifestyle changes adopted during CR once patients return to the obligations of everyday life. Our aim is to improve users' commitment to physical activity (PA) immediately after they leave the CR program by providing effective ongoing support through mHealth. This will be achieved by sending notifications on mobile devices prompting short bouts of activity. The commitment to PA is measured through weekly or monthly surveys, resulting in sparsely observed rewards. We employ reinforcement learning to determine whether and when to deliver notifications. Causal information provided by domain experts is leveraged to speed up learning. We will also discuss practical challenges in designing the learning algorithm and constructing a simulation testbed.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S75: MODERN DESIGNS FOR ADDRESSING MULTIPLICITY ISSUES IN CONFIRMATORY CLINICAL TRIALS

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Ocean Way (Mezzanine Level)

Organizer: Wenying Deng, Regeneron Pharmaceuticals
Chair: Bret Musser, Regeneron Pharmaceuticals

10:00 AM–10:20 AM Speaker: Cyrus Mehta, Cytel Inc.

Graph based adaptive group sequential designs for trials with multiple endpoints

Author(s): Cyrus Mehta, Cytel Inc.

The graph-based approach is an extremely powerful and intuitive tool for designing trials that have multiple endpoints. This method enables a study team to represent clearly its priorities for hierarchical testing of the endpoints, and for propagating the available type-1 error from rejected hypotheses to hypotheses yet to be tested. While originally developed for single stage non-adaptive designs it has recently been extended to two-stage designs that permit adaptive sample size re-estimation, dropping of hypotheses, and changes in the hierarchical testing strategy at the end of stage one. Two approaches are available for preserving the family wise error rate (FWER) in the presence of these adaptive changes; the p-value combination (PV) method, and the conditional error rate (CER) method. In this session we will present the statistical methodology underlying each approach and will compare the operating characteristics of the two methods in a large simulation experiment.

10:20 AM–10:20 AM Speaker: Siyu Li, Regeneron Pharmaceuticals

Timing and information fraction at interim analysis in event driven multi-arm studies

Author(s): Siyu Li, Regeneron Pharmaceuticals

Alpha spending function is widely used for group sequential design, which is an increasing function of the information fraction, t, at the time of the analyses. t=1 at the time of final analysis and all the alpha is spent at that time, cumulatively. For event driven multi-arm studies, it is unclear how the information fraction should be calculated and when to perform the interim/final analyses. In this study, we evaluate the commonly used methods, counting number of events in (1) paired groups that were used for sample size calculation, (2) all the treatment/control groups, (3) control group. We then evaluate the "whichever first" strategy, which is counting number of events using all the previous 3 methods and adopt the one that firstly reaches the expected number of events for the first interim analysis. The simulation results show that the "whichever first" strategy works the best on average, which has the Type I error under control, significantly shorter timing for the interim analysis when the treatment effect is much better than expected with a little decrease in power at the interim analysis and not sacrificing the overall power. Meanwhile, it keeps the blinding as needed.

10:40 AM–11:00 AM Speaker: Pranab Ghosh, Pfizer Inc.

Decision theoretic procedure for multiple testing through deep neural networks
Author(s): Pranab Ghosh, Pfizer, Inc.; Margaret Gamalo, Pfizer, Inc.

Testing multiple endpoints/doses is becoming a more common practice in recent clinical trials and graph-based testing strategy (Bretz et al. 2011) more popular because graphical representation facilitates discussion with the study team. Constructions of graph are based on clinical importance of endpoints that are part of the multiplicity strategy but quantifying the importance by assigning gain and loss function associated with each endpoint (Bauer and Bretz 2023) will allow finding the optimal graph-based strategy. This presentation will illustrate the optimization of the graph-based strategy with respect to study objectives.

11:00 AM–11:20 AM Discussant: Ming Tan, Georgetown University

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S76: ADVANCES IN COVARIATE ADJUSTMENT FOR CLINICAL TRIALS

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Platinum (Lower Level)

Organizer: Erik Bloomquist, Merck & Co., Inc.
Chair: Yu Cheng, University of Pittsburgh

10:00 AM–10:20 AM Speaker: Yue Shentu, Merck & Co., Inc.

Covariate-adjusted value-guided subgroup identification via boosting

Author(s): Jinchun Zhang, Merck & Co., Inc.; Pingye Zhang, Gilead Sciences, Inc.; Junshui Ma, Merck & Co., Inc.; Yue Shentu, Merck & Co., Inc.

It is widely recognized that treatment effects could differ across subgroups of patients. Subgroup analysis, which assesses such heterogeneity, provides valuable information in developing personalized therapies. There has been extensive research developing novel statistical methods for subgroup identification. The recent contribution is a value-guided subgroup identification method that directly maximizes treatment benefit at the subgroup level for survival outcome, rather than relying on individual treatment effect estimation. In this paper, we first completed this framework by illustrating its application to continuous and binary outcomes. More importantly, we extended the original framework to account for the prognostic effects and named this new method Covariate-Adjusted Value-guided subgroup identification via boosting (CAVboost). The original method directly used the outcome to formulate the value function for subgroup identification. Since the outcome can further be decomposed as prognostic effects and treatment effects, specifying the prognostic effects as the covariates of a model for the outcome can single out the treatment effects and improve the power to detect them across subgroups. Our proposed CAVboost was based on this key idea. It used a covariate-adjusted treatment effect estimator, instead of the outcome itself, to formulate the value function for subgroup identification. CAVboost estimates the treatment effect by using covariates to account for the prognostic effects, which mimics the idea of using covariates in an ANCOVA estimator. We showed that CAVboost could effectively improve the subgroup identification capability for both continuous and binary outcomes.

10:20 AM–10:20 AM Speaker: Dong Xi, Gilead Sciences, Inc.

Covariate adjustment and estimation of difference in proportions in randomized clinical trials
Author(s): Dong Xi, Gilead Sciences, Inc.; Jialuo Liu, Gilead Sciences, Inc.

Difference in proportions is frequently used to measure treatment effect for binary outcomes in randomized clinical trials. The estimation of difference in proportions can be assisted by adjusting for prognostic baseline covariates to enhance precision and bolster statistical power. Standardization or G-computation is a widely used method for covariate adjustment in estimating unconditional difference in proportions, because of its robustness to model misspecification. Various inference methods have been proposed to quantify the uncertainty and confidence intervals based on large-sample theories. However, their performances under small sample sizes and model misspecification have not been comprehensively evaluated. We propose an alternative approach to estimate the unconditional variance of the standardization estimator based on the robust sandwich estimator to further enhance the finite sample performance. Extensive

simulations are provided to demonstrate the performances of the proposed method, spanning a wide range of sample sizes, randomization ratios, and model misspecification. We apply the proposed method in a real data example to illustrate the practical utility.

10:40 AM–11:00 AM Speaker: Bingkai Wang, University of Michigan

Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment

Author(s): Bingkai Wang, Ryoko Susukida, Ramin Mojtabai, Masoumeh Amin-Esmaeili, Michael Rosenblum

Two commonly used methods for improving precision and power in clinical trials are stratified randomization and covariate adjustment. However, many trials do not fully capitalize on the combined precision gains from these two methods, which can lead to wasted resources in terms of sample size and trial duration. We derive consistency and asymptotic normality of model-robust estimators that combine these two methods, and show that these estimators can lead to substantial gains in precision and power. Our theorems cover a class of estimators that handle continuous, binary, and time-to-event outcomes; missing outcomes under the missing at random assumption are handled as well. For each estimator, we give a formula for a consistent variance estimator that is model-robust and that fully captures variance reductions from stratified randomization and covariate adjustment. Also, we give the first proof (to the best of our knowledge) of consistency and asymptotic normality of the Kaplan-Meier estimator under stratified randomization, and we derive its asymptotic variance. The above results also hold for the biased-coin covariate-adaptive design. We demonstrate our results using data from three trials of substance use disorder treatments, where the variance reduction due to stratified randomization and covariate adjustment ranges from 1% to 36%. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

11:00 AM–11:20 AM Discussant: Zhuqing Liu, Eli Lilly and Company

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S77: RECENT ADVANCES IN BAYESIAN INFERENCE FOR NETWORK AND TENSOR DATA

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Sound Emporium A/B (Mezzanine Level)

Organizer: Joshua Loyal, Florida State University
Chair: Joshua Loyal, Florida State University

10:00 AM–10:20 AM Speaker: Sharmistha Guha, Texas A&M University

Supervised modeling of heterogeneous networks: Investigating functional connectivity across various cognitive control tasks

Author(s): Sharmistha Guha, Texas A&M University; Ivo Dinov, University of MIchigan

> We present a novel Bayesian approach to address limitations in current methods for studying the relationship between functional connectivity across cognitive control domains and cognitive phenotypes. Our integrated framework jointly learns heterogeneous networks with vector-valued predictors, overcoming the constraints of treating each network independently in regression analysis. By assuming shared nodes across networks with varying interconnections, our method captures complex relationships while offering uncertainty quantification. Theoretical analysis demonstrates convergence to the true data-generating density, supported by empirical studies showcasing superior performance over existing approaches.

10:20 AM–10:20 AM Speaker: Michael Jauch, Florida State University

Constructing prior distributions for structured orthogonal matrices via polar expansion

Author(s): Michael Jauch, Florida State University; Marie-Christine Düker, Friedrich-Alexander University

> Statistical models for network and tensor data often include a parameter belonging to the set of orthogonal matrices. In many applications, there is reason to expect that the orthogonal matrix parameter satisfies a structural assumption such as sparsity or smoothness. In this work, we introduce an approach to constructing prior distributions for structured orthogonal matrices that leads to tractable posterior inference via parameter expanded MCMC. We draw upon results from random matrix theory to establish theoretical properties of the proposed family of prior distributions and consider applications to real data.

10:40 AM–11:00 AM Speaker: Peng Zhao, University of Delaware

Factorized fusion shrinkage for dynamic relational data

Author(s): Peng Zhao, University of Delaware; Anirban Bhattacharya, Texas A&M University; Debdeep Pati, Texas A&M University; Bani Mallick, Texas A&M University

> Modern data science applications often involve complex relational data with dynamic structures. In systems that experience regime changes, such as changes in alliances between nations after a war or air transportation networks in the wake of the COVID-19 pandemic, abrupt alterations in the relational dynamics of such data are commonly observed. To address this scenario, we propose a Factorized Fusion Shrinkage Model, which consists of a dynamic shrinkage of each decomposed factor towards a group-wise fusion structure, where shrinkage is achieved through the application of global-local shrinkage priors to successive differences in the row

vectors of the factorized matrices. The priors employed in the model preserve both the separability of clusters and the long-range properties of latent factor dynamics. Under specific conditions, we prove that the posterior distribution of the model attains the minimax optimal rate up to logarithmic factors. In terms of computation, we introduce a structured mean-field variational inference algorithm that balances optimal posterior inference with computational scalability. This framework leverages both the inter-component dependence and the temporal dependence across time. Our framework is versatile and can accommodate a wide range of models, including latent space models for networks, dynamic matrix factorization, and low-rank tensor models. The efficacy of our methodology is tested through extensive simulations and real-world data analysis.

11:00 AM–11:20 AM Speaker: Joshua Loyal, Florida State University

Fast variational inference of latent space models for dynamic networks using Bayesian p-splines
Author(s): Joshua Loyal, Florida State University

Latent space models (LSMs) are often used to analyze dynamic (time-varying) networks that evolve in continuous time. Existing approaches to Bayesian inference for these models rely on Markov chain Monte Carlo algorithms, which cannot handle modern large-scale networks. To overcome this limitation, we introduce a new prior for continuous-time LSMs based on Bayesian P-splines that allows the posterior to adapt to the dimension of the latent space and the temporal variation in each latent position. We propose a stochastic variational inference algorithm to estimate the model parameters. We use stochastic optimization to subsample both dyads and observed time points to design a fast algorithm that is linear in the number of edges in the dynamic network. Furthermore, we establish non-asymptotic error bounds for point estimates derived from the variational posterior. To our knowledge, this is the first such result for Bayesian estimators of continuous-time LSMs. Lastly, we use the method to analyze a large data set of international conflicts consisting of 4,456,095 relations from 2018 to 2022.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S78: RECENT DEVELOPMENTS IN METHODS AND APPLICATIONS OF MASTER PROTOCOLS

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Southern Ground A/B (Mezzanine Level)

Organizer: Hongtao Zhang, Merck & Co., Inc.
Chair: Heng Zhou, Merck & Co., Inc.

10:00 AM–10:20 AM Speaker: Heng Zhou, Merck & Co., Inc.

Generalized design of basket trials with p-value combination test

Author(s): Heng Zhou, Merck & Co., Inc.; Cong Chen, Merck & Co., Inc.; Linda Sun, Merck & Co., Inc.; Fang Liu, Merck & Co., Inc.

The oncology exploratory basket trial design with pruning and pooling (P&P) approach has gained increasing popularity in recent years for its simplicity and efficiency. This method was proposed based on binary endpoint, limiting its wider application. In this presentation, we will propose a generalized framework of using p-value combination test to implement pruning and pooling process in basket trials. Only p-values of any type of statistical testing from each cohort are needed for decision making, which provides great flexibility for basket trial designs with P&P approach. Numerical studies will show the trend of overall type I error and power given the p-value threshold for the combination test.

10:20 AM–10:20 AM Speaker: Yujun Wu, Morphic Therapeutic

Single sponsored platform trials: lessons learned and recommendations from a cross-industry interview

Author(s): Cindy Lu, AstraZeneca; Yujun Wu, Morphic Therapeutic

Since the concept of platform trials initially emerged in the clinical trial arena more than a decade ago, the sponsors of such a complex trial design have also been evolving. Originally platform trials offered a collaborative platform for multiple organizations and companies to develop treatment(s) for a single disease under one single protocol infrastructure. Slowly pharmaceutical companies and large research institutes realized such a trial design can be used to screen, select and prioritize their internal assets. On the other hand, special challenges arise with such single sponsored platform trials. Aiming to understand the challenges and solutions for such trials design and implementations, we hand-picked 10 pharmaceutical companies that are routinely conducting such trials and interviewed their cross-functional team. This presentation will summarize the ongoing effort of the interview and share the learnings and recommendations from those teams.

10:40 AM–11:00 AM Speaker: Robert Beckman, Georgetown University

Generalized randomized confirmatory basket trials: Efficiency, type I error control, and a simulated application

Author(s): Daphne Guinn, Georgetown University Medical Center; Linchen He, Novartis Pharmaceuticals; Sathvik Narayana, Union College; Valeriy Korostyshevskiy, Georgetown University Medical Center; Robert Beckman, Georgetown University Medical Center

Basket trials pool indications sharing molecular pathophysiology, improving development efficiency. Generalized application of basket trials in the confirmatory phase, where most development resources are expended, could have a marked impact on efficiency. To date, basket trials have been confirmatory only for exceptional therapies and in end-stage patients. Our previous randomized basket design may be generally suitable in the confirmatory phase, maintains high power even with modest effect sizes, and provides nearly k-fold increased efficiency for k indications, but controls false positives for the pooled result only. Since family-wise error rate by indications (FWER) may sometimes be required, we now simulate a variant of this basket design controlling FWER at 0.025k, the total FWER of k separate randomized trials. We also report the simulation of a specific application in autoimmune diseases, based on both the literature and on real world data from off-label use in the Georgetown-Medstar database. Even under FWER control, randomized confirmatory basket trials substantially improve development efficiency, and RWD is helpful to inform design simulations.

11:00 AM–11:20 AM Discussant: Li Wang, AbbVie Inc.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S79: IISA SESSION: ON SOME RECENT ADVANCES IN BAYESIAN STATISTICS

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Blackbird B (Mezzanine Level)

Organizer: Bodhisattva Sen, Columbia University
Chair: Anindya Bhadra, Purdue University
IISA = International Indian Statistical Association

10:00 AM–10:20 AM Speaker: Anirban Bhattacharya, Texas A&M University

Hybrid approximation to marginal likelihood

Author(s): Eric Chuu, Texas A&M University; Yabo Niu, University of Houston; Anirban Bhattacharya, Texas A&M University; Debdeep Pati, Texas A&M University

We consider the estimation of the marginal likelihood in Bayesian statistics, with primary emphasis on Gaussian graphical models, where the intractability of the marginal likelihood in high dimensions is a frequently researched problem. We propose a general algorithm that can be widely applied to a variety of problem settings and excels particularly when dealing with near log-concave posteriors. Our method builds upon a previously posited algorithm that uses MCMC samples to partition the parameter space and forms piecewise constant approximations over these partition sets as a means of estimating the normalizing constant. In this paper, we refine the aforementioned local approximations by taking advantage of the shape of the target distribution and leveraging an expectation propagation algorithm to approximate Gaussian integrals over rectangular polytopes. Our numerical experiments show the versatility and accuracy of the proposed estimator, even as the parameter space increases in dimension and becomes more complicated.

10:20 AM–10:20 AM Speaker: Jyotishka Datta, Virginia Tech

Quantile importance sampling and inverse probability weighting

Author(s): Jyotishka Datta, Virginia Tech; Nicholas Polson, University of Chicago

In Bayesian inference, evidence estimation is a fundamental challenge as it is important for various purposes, including model selection. The existing strategies for evidence estimation are classified into four categories: deterministic approximation, density estimation, importance sampling, and vertical representation (Llorente, 2020). In this talk, we show that the Riemann sum estimator due to Yakowitz (1978) can be used in the context of nested sampling Skilling (2006) to achieve a $O(n^{-4})$ rate of convergence, faster than the usual Ergodic Central Limit Theorem, under certain regularity conditions. If time permits, we will discuss a "weak paradox" related to evidence estimation involving the inverse probability weight (IPW) estimators. IPW estimators include the popular Horwitz–Thompson and Hajek estimators, and are used routinely in survey sampling, causal inference and evidence estimation for Bayesian computation.

10:40 AM–11:00 AM Speaker: Debdeep Pati, Texas A&M University

Blocked Gibbs sampler for hierarchical Dirichlet processes

Author(s): Snigdha Das, Texas A&M University; Yabo Niu, University of Houston; Yang Ni, Texas A&M University; Bani Mallick, Texas A&M University; Debdeep Pati, Texas A&M University

Posterior computation in hierarchical Dirichlet process (HDP) mixture models is an active area of research in nonparametric Bayes inference of grouped data. Existing literature almost exclusively focuses on the Chinese restaurant franchise (CRF) analogy of the marginal distribution of the parameters, which can mix poorly and is known to have a linear complexity with the sample size. A recently developed slice sampler allows for efficient blocked updates of the parameters, but is shown to be statistically unstable in our article. We develop a blocked Gibbs sampler to sample from the posterior distribution of HDP, which produces statistically stable results, is highly scalable with respect to sample size, and is shown to have good mixing. The heart of the construction is to endow the shared concentration parameter with an appropriately chosen gamma prior that allows us to break the dependence of the shared mixing proportions and permits independent updates of certain log-concave random variables in a block. En route, we develop an efficient rejection sampler for these random variables leveraging piece-wise tangent-line approximations.

11:00 AM–11:20 AM Speaker: Anindya Bhadra, Purdue University

Likelihood based inference in exponential family graphical models with intractable normalizing constants

Author(s): Anindya Bhadra, Purdue University

Probabilistic graphical models that encode an underlying Markov random field are fundamental building blocks of generative modeling to learn latent representations in modern multivariate data sets with complex dependency structures. Among these, the exponential family graphical models are especially popular given their fairly well-understood statistical properties and computational scalability to high-dimensional data based on pseudo-likelihood methods. These models have been successfully applied in many fields, such as the Ising model in statistical physics and count graphical models in genomics. Another strand of models allows some nodes to be latent, so as to allow the marginal distribution of the observable nodes to depart from exponential family to capture more complex dependence. These approaches form the basis of generative models in artificial intelligence, such as the Boltzmann machines and their restricted versions. A fundamental barrier to likelihood-based (i.e., both maximum likelihood and fully Bayesian) inference in both fully and partially observed cases is the intractability of the likelihood. The usual workaround is via adopting pseudolikelihood based approaches, following the pioneering work of Besag (1974). The goal of this paper is to demonstrate that full likelihood based analysis of these models is feasible in a computationally efficient manner. The chief innovation lies in using a technique of Geyer (1991) to estimate the intractable normalizing constant, as well as its gradient, for intractable graphical models. Extensive numerical results, supporting theory and comparisons with pseudolikelihood based approaches demonstrate the applicability of the proposed method.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S80: RECENT ADVANCEMENTS ON CAUSAL MEDIATION ANALYSIS

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Gold (Lower Level)

Organizer: Yinqiu He, University of Wisconsin-Madison
Chair: Yinqiu He, University of Wisconsin-Madison

10:00 AM–10:20 AM Speaker: Wei Hao, University of Michigan

A class of directed acyclic graphs with mixed data types in mediation analysis

Author(s): Wei Hao, University of Michigan; Canyi Chen, University of Michigan; Peter Song, University of Michigan

We propose a unified class of generalized structural equation models (GSEMs) with data of mixed types in mediation analysis, including continuous, categorical, and count variables. Such models extend substantially the classical linear structural equation model to accommodate many data types arising from the application of mediation analysis. Invoking the hierarchical modeling approach, we specify GSEMs by a copula joint distribution of outcome variable, mediator, and exposure variable, in which marginal distributions are built upon generalized linear models (GLMs) with confounding factors. We discuss the identifiability conditions for the causal mediation effects in the counterfactual paradigm as well as the issue of mediation leakage, and develop an asymptotically efficient profile maximum likelihood estimation and inference for two key mediation estimands, natural direct effect and natural indirect effect, in different scenarios of mixed data types. The proposed new methodology is illustrated by a motivating epidemiological study that investigates whether the tempo of reaching infancy BMI peak (delay or on time), an important early life growth milestone, mediates the association between prenatal exposure to phthalates and pubertal health outcomes.

10:20 AM–10:20 AM Speaker: Chunlin Li, Iowa State University

Quantifying the global mediation effect for nonsparse high-dimensional genomics mediators

Author(s): Tianzhong Yang, University of Minnesota; Zhiyu Kang, University of Minnesota; Chunlin Li, Iowa State University

While many existing epidemiological studies have examined associations between alcohol and cardiovascular outcomes, less has been done to explore causal biological pathways and mechanisms of the observed associations at the molecular level. To investigate this relationship, we propose a new causal measure to quantify the mediating role of molecular phenotypes, such as DNA methylation, in bridging alcohol intake and cardiovascular outcomes. The challenge of estimating this measure is two-fold. First, since alcohol consumption is associated with genome-wide changes at the molecular level, it is biologically plausible that many omics mediators with weak but collectively considerable effects are involved in the pathway; however, existing methods are plagued by inconsistency in the presence of non-sparse mediators. To address this issue, we develop a method to estimate the proposed measure in such situations consistently. Second, many epidemiological studies use case-control sampling, which introduces ascertainment bias in mediation analysis. To correct this bias, we propose a method of moment motivated by heritability estimation. Finally, a significant challenge in this research is the potential for residual confounding in observational studies, which can seriously com-

promise the validity of scientific findings. We will briefly discuss the approach to correct the confounding bias.

10:40 AM–11:00 AM Speaker: Yinqiu He, University of Wisconsin-Madison

Adaptive bootstrap tests for composite null hypotheses in the mediation pathway analysis
Author(s): Yinqiu He, University of Wisconsin-Madison; Peter X.K. Song, University of Michigan-Ann Arbor; Gongjun Xu, University of Michigan-Ann Arbor

Mediation analysis aims to assess if, and how, a certain exposure influences an out- come of interest through intermediate variables. This problem has recently gained a surge of attention due to the tremendous need for such analyses in scientific fields. Testing for the mediation effect is greatly challenged by the fact that the underlying null hypothesis (i.e., the absence of mediation effects) is composite. Most existing mediation tests are overly conservative and thus underpowered. To overcome this significant methodological hurdle, we develop an adaptive boot- strap testing framework that can accommodate different types of composite null hypotheses in the mediation pathway analysis. Applied to the product of coefficients (PoC) test and the joint significance (JS) test, our adaptive testing procedures provide type I error control under the composite null, resulting in much improved statistical power compared to existing tests. Both theoretical properties and numerical examples of the proposed methodology are discussed.

11:00 AM–11:20 AM Speaker: Canyi Chen, University of Michigan

Quantile mediation analytics
Author(s): Canyi Chen, University of Michigan; Yinqiu He, University of Wisconsin; Huixia J. Wang, The George Washington University; Gongjun Xu, University of Michigan; Peter X.-K. Song, University of Michigan

Mediation analysis is used to examine if and how an intermediate variable mediates the influence of an exposure variable on an outcome of interest. Quantiles, rather than the mean, of an outcome are scientifically relevant to the comparison among specific subgroups in practical studies. Albeit some empirical studies are available in the literature, there lacks a thorough theoretical investigation of quantile-based mediation analysis, which hinders practitioners from using such methods to answer important scientific questions. To address this significant technical gap, in this paper, we develop a quantile mediation analysis methodology to facilitate identifying, estimating, and testing quantile mediation effects under a hypothesized directed acyclic graph. We establish two key estimands, quantile natural direct effect and quantile natural indirect effect in the counterfactual framework, both of which have closed-form expressions. To overcome the issue that the null hypothesis of no mediation effect is composite, we propose an adaptive bootstrap method that is shown theoretically and numerically to achieve a proper type I error control. We illustrate the proposed quantile mediation analysis methodology through both extensive simulation experiments and real-world data in that we investigate the mediation effect of lipidomic biomarkers for the influence of exposure to phthalates on early childhood obesity defined clinically by 95% percentile of their body mass index.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

S81: ADVANCES IN COMPUTATIONAL ALGORITHMS IN COMPLEX BIOMEDICAL END-POINTS AND SMALL AREA ESTIMATION

Wednesday, June 19, 2024
10:00 AM–11:30 AM, Melody (Lobby Level)

Organizer: Priyam Das, Virginia Commonwealth University
Chair: Debamita Kundu, University of Virginia

10:00 AM–10:20 AM Speaker: Xuan Wang, University of Utah

SurvMaximin: Robust federated approach to transporting survival risk prediction models

Author(s): Xuan Wang, University of Utah

For multi-center heterogeneous Real-World Data (RWD) with time-to-event outcomes and high-dimensional features, we propose the SurvMaximin algorithm to estimate Cox model feature coefficients for a target population by borrowing summary information from a set of health care centers without sharing patient-level information. For each of the centers from which we want to borrow information to improve the prediction performance for the target population, a penalized Cox model is fitted to estimate feature coefficients for the center. Using estimated feature coefficients and the covariance matrix of the target population, we then obtain a SurvMaximin estimated set of feature coefficients for the target population. The target population can be an entire cohort comprised of all centers, corresponding to federated learning, or a single center, corresponding to transfer learning. Simulation studies and a real-world international electronic health records application study, with 15 participating health care centers across three countries (France, Germany, and the U.S.), show that the proposed SurvMaximin algorithm achieves comparable or higher accuracy compared with the estimator using only the information of the target site and other existing methods. The SurvMaximin estimator is robust to variations in sample sizes and estimated feature coefficients between centers, which amounts to significantly improved estimates for target sites with fewer observations. The SurvMaximin method is well suited for both federated and transfer learning in the high-dimensional survival analysis setting. SurvMaximin only requires a one-time summary information exchange from participating centers. Estimated regression vectors can be very heterogeneous. SurvMaximin provides robust Cox feature coefficient estimates without outcome information in the target population and is privacy-preserving.

10:20 AM–10:20 AM Speaker: Priyam Das, Virginia Commonwealth University

Clustering sequence data with mixture Markov chains with covariates using Multiple Simplex Constrained Optimization Routine (MSiCOR)

Author(s): Priyam Das, Virginia Commonwealth University; Deborshee Sen, Google India; Debsurya De, Johns Hopkins University; Jue Hou, University of Minnesota; Zahra Abad, University of Toronto; Nicole Kim, Harvard T.H. Chan School of Public Health; Zongqi Xia, University of Pittsburgh; Tianxi Cai, Harvard T.H. Chan School of Public Health

Mixture Markov Model (MMM) is a widely used tool to cluster sequences of events coming from a finite state-space. However, the MMM likelihood being multi-modal, the challenge remains in its maximization. Although the Expectation-Maximization (EM) algorithm remains one of the most popular ways to estimate the MMM parameters, however, convergence of the

EM algorithm is not always guaranteed. Given the computational challenges in maximizing the mixture likelihood on the constrained parameter space, we develop a pattern search-based global optimization technique which can optimize any objective function on a collection of simplexes, which is eventually used to maximize MMM likelihood. This is shown to outperform other related global optimization techniques. In simulation experiments, the proposed method is shown to outperform the expectation-maximization (EM) algorithm in the context of MMM estimation performance. The proposed method is applied to cluster Multiple sclerosis (MS) patients based on their treatment sequences of disease-modifying therapies (DMTs). We also propose a novel method to cluster people with MS based on DMT prescriptions and associated clinical features (covariates) using MMM with covariates. Based on the analysis, we divided MS patients into three clusters. Further cluster-specific summaries of relevant covariates indicate patient differences among the clusters.

10:40 AM–11:00 AM Speaker: Guoyi Zhang, University of New Mexico

Unit level small area estimation using support vector machine
Author(s): Guoyi Zhang, University of New Mexico; John Pleis, National Center for Health Statistics

The BHF unit level model (Battese, Carter & Fuller, 1988) is popularly used in small area estimation (SAE). The BHF model assumes a linear relationship between the response and predictor variables plus a random effect term for each area. Although the dependency of the observations within each area has been taken accounted for, the linear relationship may not be appropriate. In this research, we extend the BHF model to a more general semi-parametric model for small area estimation (SP-SAEunit), in which the nonparametric component is estimated by machine learning tools such as support vector machine. A back fitting algorithm is proposed to solve the SP-SAEunit problem. Simulation studies show that the proposed SP-SAEunit model significantly outperforms the direct estimation and BHF models when linear relationship has been violated.

11:00 AM–11:20 AM Speaker: Dipankar Bandyopadhyay, Virginia Commonwealth University

Exploring EM accelerations for longitudinal EHR data with matrix-variate non-Gaussian responses
Author(s): Qingyang Liu, Sanvesh Srivastava, Dipankar Bandyopadhyay

Matrix-variate (MV) continuous responses abound in biomedical studies, such as in longitudinal periodontal disease (PD) monitoring in electronic health records (EHR). However, the analysis of this data poses various challenges. For example, the responses constituting the MV structure (such as, pocket depth and clinical attachment level for PD data) can be Non-Gaussian, with considerable skewness and heavy tails, where implementing a standard matrix-normal regression modeling can be suboptimal. Furthermore, the size of the dataset (due to large n subjects) can easily complicate an EM-type estimation setup, rendering it unscalable. To overcome these, in this talk, we cast this regression problem into a MV skew-t regression framework, allowing the dimension of the matrices for each subject to vary, thus mimicking an EHR setup. Additionally, for estimation, we propose two computationally elegant schemes: (a) a general divide-and-combine EM approach, and (b) an asynchronous and distributed data augmentation EM framework via distributed computing. The novelty of the asynchronous scheme lies in guaranteeing that the distance between the target and estimated densities become negligible, provided that the data augmentation algorithm runs sufficiently long, and thus overcome

the key drawbacks of the usual divide-and-combine algorithms. Theoretical guarantees for the asymptotic optimality of these algorithms under mild conditions will be presented. Furthermore, findings from simulation studies and application to a real PD EHR database highlighting the relative advantages of our proposal, both in terms of adequacy of model fit and scalability, will be discussed.

11:20 AM–11:30 AM **Q&A and Floor Discussion**

See you at
the 2025 ICSA Applied Statistics Symposium!

University of Connecticut

Qiqi Deng and Xiaojing Wang, co-chairs

https://www.icsa.org/

©2024 International Chinese Statistical Association

version 1.2