

IDWSDS 2026



October 6, 2026

International Day of Women in Statistics and Data Science

Creating Global Connections

2026 INTERNATIONAL DAY OF WOMEN IN STATISTICS AND DATA SCIENCE

CONFERENCE PROGRAM AND ABSTRACTS

OCTOBER 6TH, 2026

A VIRTUAL 24 HOUR CONFERENCE

[HTTPS://IDWSDS.ORG](https://idwsds.org)

ORGANIZED BY

CAUCUS FOR WOMEN IN STATISTICS AND DATA SCIENCE

HOSTED BY

CAUCUS FOR WOMEN IN STATISTICS AND DATA SCIENCE VIA ZOOM

© 2026

CAUCUS FOR WOMEN IN STATISTICS AND DATA SCIENCE

[HTTPS://WWW.CWSTAT.ORG](https://www.cwstat.org)



Welcome



We are delighted to welcome you to the 2026 International Workshop on Statistical Data Science (IDWSDS), where this year's theme is "Creating Global Connections." As researchers, practitioners, educators, and students, our work is strengthened by the connections we make across disciplines, institutions, countries, and perspectives. This year's program celebrates the power of bringing people and ideas together to build new collaborations, share knowledge, and advance statistical data science across borders.

Whether you are joining us for the first time or returning to IDWSDS, we are excited to have you with us. This year's program brings together a diverse community and offers opportunities to learn from one another, discover new perspectives, and build connections that extend beyond the conference.

Conference App – Sched

We're continuing to use Sched as your digital conference guide. All session information, speaker bios, schedules, and conference updates are available through the app. Use Sched to build your personalized agenda, keep up with changes, and discover other attendees and speakers. We encourage you to explore the platform and use it as a way to connect with members of our global community throughout the conference.

Connecting Across Ideas and Borders

Throughout the program, you'll find talks and sessions that highlight the many ways statistical data science brings people and disciplines together. From new research and methodologies to practical applications and shared challenges, we hope the program sparks conversations that continue long after a session ends.

Take the opportunity to introduce yourself, ask questions, share your experiences, and seek out perspectives beyond your usual networks. Some of the most valuable connections at a conference happen between sessions, over a conversation, or through an unexpected exchange of ideas.

Acknowledging Our Sponsors

We are proud to recognize and thank our generous sponsors for their support of IDWSDS:

Platinum Level

- CANSII
- Institute of Mathematical Statistics
- Statistical Information Institute for Consulting and Analysis & Japan Statistical Society

Gold Level

- MRC Biostatistics Unit - University of Cambridge

Silver Level

- StataCorp
- Bernoullis Society for Mathematical Statistics and Probability

Bronze Level

- CRC Press

Their support helps make this conference possible and strengthens the community that IDWSDS brings together. We encourage you to learn more about our sponsors and engage with their sessions throughout the program.

We hope you leave IDWSDS 2026 with new ideas, new collaborations, and new connections that continue well beyond the conference. Welcome, and enjoy the conference!

Cynthia Bland, Jessica Kohlschmidt, and Samantha Seals
Co-chairs, 2026 International Day of Women in Statistics and Data Science Organizing Committee



Organizing Committee

CO-CHAIR



CYTHIA BLAND

RTI and CWS

Cynthia Bland is the Past President of the Caucus for Women in Statistics and Data Science. This year she has chaired the nominating committee and an award committee. Professionally she is a survey statistician and directs the Center for Official Statistics at RTI International, a non-profit research firm specializing in government contracting. In her two decades at RTI, she has designed and analyzed national surveys, with an emphasis on patient experience of care surveys. Cynthia leads a center of more than 70 statisticians as they innovate for clients and grow in their own careers. Cynthia serves as the Deputy Editor in Chief of RTI Press, an open-access, peer-reviewed journal, has been active with the NC Chapter of the American Statistical Association, and is a lecturer in statistics for the Kenan-Flagler School of Business at the University of North Carolina at Chapel Hill. Outside of the statistical world, she serves on the board of a local credit union and enjoys ballet.

<https://www.researchgate.net/profile/Ligia-Rodrigues-2>


LÍGIA HENRIQUES-RODRIGUES

University of Évora and CIMA

Lígia Henriques-Rodrigues (LHR) is an Assistant Professor in the Department of Mathematics at the School of Science and Technology, University of Évora. LHR completed a postdoctoral fellowship in Statistics of Extremes at the Faculty of Sciences, University of Lisbon, and earned her Ph.D. in Probability and Statistics from the same university. Her primary research focuses on the statistics of extremes and computational statistics, particularly in developing new classes of semi-parametric reduced-bias estimators for extreme event parameters. The applications of her research span environmental and financial sciences. Beyond her academic pursuits, LHR has been a board member of the Portuguese Statistical Society (SPE) since January 2024 and is a member of the executive committee for the Master's Degree in Statistical Modeling and Data Analysis at the University of Évora.

CO-CHAIR



JESSICA KOHLSCHMIDT

The Ohio State University and CWS

Dr. Jessica K. Kohlschmidt is a distinguished biostatistician, educator, and leader in the statistical community. A first-generation college student, her early passion for math and teaching evolved into a dedicated career in statistics after discovering its real-world applications.

She has built a multifaceted career that combines teaching, high-impact clinical research, and significant leadership roles in professional organizations. For 12 years, she applied her expertise to leukemia outcomes research at The Ohio State University, and she has been a college educator since 1999.

Dr. Kohlschmidt is deeply committed to mentorship and outreach, particularly encouraging young women to pursue statistics, and holds key leadership positions in several statistical organizations, including the American Statistical Association (ASA).

<https://fisher.osu.edu/people/kohlschmidt.1>


ALTEA LORENZO-ARRIBAS

Biomathematics and Statistics Scotland (BioSS)

Altea is a socio-economic statistician at Biomathematics and Statistics Scotland (BioSS) working in collaboration with researchers at the Scottish Environment, Food and Agriculture Research Institutions. She is an elected council member of the Royal Statistical Society, member of the Society's AI Task Force, chair of the Celebrating Diversity Special Interest Group and the secretary of the History of Statistics Section, as well as a member of the Women Committee of the Spanish Society of Statistics and Operations Research (SEIO), and the Spanish Biostatistics Network (Biostatnet).

<https://www.bioss.ac.uk/people/altea>


VANDA M. LOURENÇO

NOVA University of Lisbon

Vanda M. Lourenço (VML) is an Assistant Professor in the Department of Mathematics at the NOVA School of Science and Technology, part of the NOVA University of Lisbon. She earned her PhD in Statistics and Stochastic Processes from Instituto Superior Técnico at the University of Lisbon. Her primary research focuses on robust, non-parametric, and computational statistics, particularly in genetic and genomic association studies and the prediction of quantitative traits. While her expertise has been predominantly applied to challenges in plant breeding, she is also keen on extending these methodologies to animal and human studies. In addition to her academic work, VML actively contributes to professional and statistical communities. She served as President of the Supervisory Board of the Portuguese Statistical Society (SPE) from 2021 to 2024 and currently represents Portugal in the Caucus for Women in Statistics (CWS), where she is also a member of the Nominations Committee. Additionally, she is an associate editor for the Journal of Data Science, Statistics, and Visualization (JDSSV), a publication of the ISI International Association for Statistical Computing. Beyond her formal roles, she is involved in various other initiatives within and outside statistical societies.



Organizing Committee

continued...



CO-CHAIR

DR. SAMANTHA SEALS

University of West Florida

<http://sealslab.github.io>

Dr. Samantha Seals is an Associate Professor of Mathematics and Statistics at the University of West Florida, where she teaches applied statistics and data science. Her research focuses on STEM student outcomes and the application of statistical modeling to ecology and environmental science problems. As a member of and volunteer for the Caucus for Women in Statistics (CWS), Dr. Seals is committed to advancing opportunities for women in statistics and data science and celebrating their contributions to the field. Her work reflects a dedication to mentoring the next generation of statisticians while applying rigorous quantitative methods to address real-world challenges.



JESSICA STOCKDALE

Simon Fraser University

Dr. Jessica Stockdale is an Assistant Professor in the Department of Mathematics at Simon Fraser University. She is an embedded researcher with Vancouver Coastal Health, where she co-leads an Embedded Research Program in mathematical modelling and operations research. Before joining SFU as a postdoctoral fellow in 2018, Jessica completed her MMath and PhD at the University of Nottingham's School of Mathematical Sciences, with a doctoral thesis on computational methods for stochastic epidemic models. Her research uses mathematical modelling, statistics and data science to advance evidence-informed, equitable health care delivery and population health interventions, with focuses on primary and community care and public health.



HUNNA WATSON

University of North Carolina at Chapel Hill

<https://www.linkedin.com/in/hunna-watson-52030352/>

Dr. Hunna Watson is a Research Associate Professor in the School of Medicine at the University of North Carolina at Chapel Hill, where she serves as a biostatistician in the Center of Excellence for Eating Disorders. She also holds Adjunct Research Fellow appointments in the School of Medicine at The University of Western Australia and in the Discipline of Psychology at Curtin University in Australia. Specializing in the field of eating disorders, Dr. Watson has authored over 120 peer-reviewed scientific articles and book chapters, focusing on critical areas such as diagnostic nosology, prevention and treatment, molecular genetics, and randomized clinical trials. Her research has gained global recognition through lectures, papers, and workshops presented worldwide. Dr. Watson serves on the editorial board of the International Journal of Eating Disorders and is a member of the Alumni Group of the National Eating Disorders Collaboration Steering Committee in Australia. Dr. Watson completed her PhD in Clinical Psychology at Curtin University and holds Master's degrees in Biostatistics from the University of Melbourne and Clinical Psychology from Curtin University.



Keynote Sessions



Session Info

K1

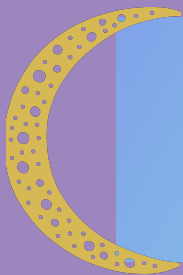
KEYNOTE SESSION

October 6th, 1:00 - 2:00 UTC

A Statistician in the Loop: Building a Career in Trustworthy AI

USE SCHED TO ACCESS THE ZOOM ROOM

Statistics provides tools for answering questions that are central to building trustworthy AI: What works? What fails? How certain are we? In this talk, I will discuss my path from studying statistics to conducting research as a data scientist and now working full time in AI safety. I will share examples of how statistical thinking can support this work, including deciding what to measure, uncovering failure modes, quantifying uncertainty, and connecting evidence to real decisions. I hope to show that many paths lead into this field and that statisticians and data scientists have important skills to bring to the challenge of building trustworthy AI.

ORGANIZER: IDWSDS Program Committee**CHAIR:** Burcu Eke Rubini, University of New Hampshire

Speaker Bio

K1



Emily Hadley

OpenAI, Quantitative Threat Forecasting Analyst

Emily Hadley is a Quantitative Threat Forecasting Analyst at OpenAI, where she works to identify, understand, and forecast risks related to AI. Previously, she was a Security Researcher on the Microsoft AI Red Team and a Senior Research Data Scientist at RTI International. Emily holds an MS in Analytics from NC State University and a BS in Statistics with a second major in Public Policy from Duke University.



Session Info

K2

KEYNOTE SESSION

October 6th, 6:00 - 7:00 UTC

Bayes in Practice: A Bayesian Cancer Atlas

USE SCHED TO ACCESS THE ZOOM ROOM

From its earliest beginnings, Bayesian statistics has been a synthesis of theory, methodology, computation and application. In this presentation, I will reflect on the development of the award-winning Australian Cancer Atlas (<https://atlas.cancer.org.au/>) and spotlight its Bayesian foundations. I will highlight some of the challenges and proposed solutions to modelling and visualisation of an awkward spatial geography, “filling in” missing covariates, and communicating uncertainty. I will also touch on new research directions inspired by the Atlas: new spatio-temporal models, spatial vulnerability indices, meta-analysis transfer learning, distributed AI and responsible data science. Importantly, these methodological discussions will be complemented by reflections on the impact of the work for patients, health practitioners, cancer support groups and government agencies. Bayesian statistics really can make a difference!

Selected References

Baade P, K Mengersen [2024] Building HOPE through the Australian Cancer Atlas Insight+ MJA 35 //insightplus.mja.com.au/2024/35

J Bon, A Bretherton, K Buchhorn, S Cramb, C Drovandi, C Hassan, A Jenner, H Mayfield, J. McGree, K Mengersen, A Price, R Salomone, E Santos-Fernandez, J Vercelloni & X Wang, [2023] Being Bayesian in the 2020s: opportunities and challenges in the practice of modern applied Bayesian statistics. *Philosophical Transactions. Series A, Mathematical, physical, and engineering sciences*, 381(2247), Article number: 20220156.

Bretherton A, Bon J, Warne D, Mengersen K, Drovandi C, [2026] A Principled Approach to Bayesian Transfer Learning, *Bayesian Analysis*. To appear.

Cramb SM, K Mengersen, and PD Baade. [2011] Developing the atlas of cancer in Queensland: methodological issues. *International Journal of Health Geographics*, 10, p 1-11, 2011

Goodwin S, T Saunders, J Aitken, P Baade, U Chandrasiri, D Cook, S Cramb, E Duncan, S Kobakian, J. Roberts, K. Mengersen, [2024] Designing the Australian Cancer Atlas: visualizing geostatistical model uncertainty for multiple audiences. *JAMIA. Journal of the American Medical Informatics Association*, //doi.org/10.1093/jamia/ocae212

Hassan, C [2024] Structured Models and Algorithms for Sensitive Data. PhD Thesis, Queensland University of Technology, Australia.

Hogg J, J Cameron, S Cramb, P Baade, K Mengersen [2024] A Two-stage Bayesian Small Area Estimation Approach for Proportions. *International Statistical Review*. V92 I3 455482.

Hyland-Wood B, Snoswell A, Sandeep R, Chun O, Perrin D, Fieft E, Price A, Mengersen K (2024) Response to proposals paper on introducing mandatory guardrails for AI in highrisk settings. *Analysis and Policy Observatory*, 2024/10/4

Jahan F, Duncan E, Cramb S, Baade P, Mengersen K, [2020] Multivariate Bayesian metaanalysis: Joint modelling of multiple cancer types using summary statistics, *International Journal of Health Geographics*, 19 (1)

Leontyeva Y, Y Huang, S Cramb, J Cameron, P Baade, K Mengersen, et al. [2025] Bayesian Spatial Relative Survival Model to Estimate the Loss in Life Expectancy and Crude Probability of Death for Cancer Patients, *Statistics in Medicine* 44 (3-4), e10287

Price A, M Rigby, P Fiévez, K Mengersen [2025] A spatial vulnerability index for environmental health *Ecological Indicators* Vol 178, September 2025, 113793

ORGANIZER: IDWSDS Program Committee

CHAIR: Altea Lorenzo-Arribas, Biomathematics and Statistics Scotland (BioSS)



Speaker Bio

K2



Kerrie Mengersen

Queensland University of Technology, Emeritus Professor

Kerrie Mengersen is an Emeritus Distinguished Professor at QUT and an internationally recognised leader in statistics, data science, and artificial intelligence. She holds a BA (Hons) in Mathematical Statistics and Computing and a PhD in Mathematical Statistics from the University of New England, Australia.

A leading expert in Bayesian statistics, Kerrie has spent more than three decades developing innovative statistical and machine learning methods to address complex challenges in health, medicine, the environment, agriculture, and industry. Her work has helped shape the growth of data science in Australia and has demonstrated the power of data and evidence-based decision-making to create real-world impact.

Kerrie is a Fellow of the Australian Academy of Science, the Academy of Social Sciences in Australia, and the Queensland Academy of Arts and Sciences. She is also a passionate mentor and advocate for the next generation of data scientists, particularly women in STEM, and continues to champion greater diversity, inclusion, and excellence across the statistical and data science professions.



Session Info

K3

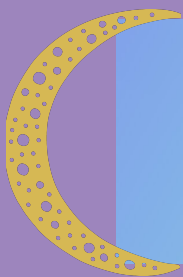
KEYNOTE SESSION

October 6th, 11:00 - 12:00 UTC

Forcing a way into the Black Box

USE SCHED TO ACCESS THE ZOOM ROOM

Modern statistical computing allows us to perform sophisticated inference through increasingly complex algorithms and software, often without fully understanding what happens inside the black box. Using Markov chain Monte Carlo as a starting point, I will discuss the role of convergence, mixing, Monte Carlo error, and asymptotic variance in assessing the reliability of computational inference. I will also connect this perspective to teaching and student mentorship, where I believe our role is to cultivate independent statistical thinking rather than simply transmit established methods. Across these themes, I will argue for the importance of understanding what lies beneath our computational tools—and knowing when we can, and cannot, trust them.

ORGANIZER: IDWSDS Program Committee**CHAIR:** Cynthia Bland, CWS Past Past President; RTI,

Speaker Bio

K3



Dootika Vats

IIT Kanpur, Department of Statistics and Data Science

Dootika Vats is an Associate Professor in the Department of Statistics and Data Science at the Indian Institute of Technology Kanpur. Her research focuses on Markov chain Monte Carlo, output analysis for stochastic simulation, and stochastic optimization. She received her PhD from the University of Minnesota, Twin Cities, and was an NSF Postdoctoral Fellow at the University of Warwick. Her work has been recognized through several teaching awards and, most recently, the 2026 COPSS Emerging Leader Award. She is also passionate about teaching and mentorship and is committed to fostering independent statistical thinking in the next generation of statisticians.



Session Info

K4

KEYNOTE SESSION

October 6th, 16:00 - 17:00 UTC

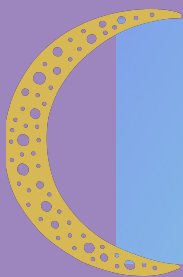
Open Science - Opportunities and Challenges

USE SCHED TO ACCESS THE ZOOM ROOM

Women statisticians and data scientists – especially those working in isolation in developing and under-resourced countries, face many barriers in their career development. It is true that “the best thing about being a statistician is that you get to play in everybody’s backyard” (John Tukey), but this also means that statisticians often work in interdisciplinary settings with limited access to mentorship or to fellow statisticians for consultation.

In this talk, I will share my experiences as an isolated applied statistician “playing in different backyards” (biomedicine, public administration, informatics) and across diverse settings (research institute, public administration, private consulting, university). Throughout my career, I have both promoted and relied on openness in science and ICT: open-source software, open textbooks, open standards, open (government and research) data, MOOCs, and open access research publications.

I will discuss the opportunities and challenges that open science presents for statistics and data science, and highlight importance of professional societies for professional development and community building.

ORGANIZER: IDWSDS Program Committee**CHAIR:** Hunna Watson, University of North Carolina at Chapel Hill



Diana Šimić

University of Zagreb Faculty of Organization and Informatics, Retired Full Professor

Diana Šimić has retired, but continues to work as a professor of statistics and research methods at the University of Zagreb Faculty of Organization and Informatics, and the International University Libertas Faculty of Health Sciences. She holds a degree in mathematical statistics and informatics engineering—one of the earliest study programs that would now be recognized as data science—and a Ph.D. in public health.

Over the course of her career, she has worked as a computer programmer (today's "developer"), applied statistician, government official, consultant in e-business and e-government, and university professor. As a government official, she successfully championed the inclusion of open access to scientific publications in the official documents of the World Summit on the Information Society (2003-2005). She has taken part in several national open science initiatives, and emphasizes the importance of open access to scientific information in her teaching of research methodology.

Diana is a founding member of the Croatian Biometric Society (recently accepted as the Croatian region of the International Biometric Society), where she served as president, general secretary, and council member. She is also a member of the Croatian Statistical Association, the American Statistical Association, and the Association for Computing Machinery.



Session Info

K5

KEYNOTE SESSION

October 6th, 19:00 - 20:00 UTC

Building Bridges That Outlive Us

USE SCHED TO ACCESS THE ZOOM ROOM

In a world where technology has made it easier than ever to connect across borders, meaningful connections are not created simply by proximity, platforms, or professional networks. They are built through trust, service, shared purpose, and a willingness to create opportunities for others.

In this keynote, I will reflect on my educational and professional journey and the unexpected role that volunteering and service have played in shaping my career and leadership. While formal employment provided opportunities to develop technical expertise and professional experience, many of the experiences that most profoundly influenced my growth came through opportunities to serve within professional associations, communities, and initiatives focused on developing people and strengthening the statistics and data science ecosystem.

These experiences taught me that leadership is not simply about the positions we occupy, but about the connections we create and the pathways we make possible for others. A conversation can become an opportunity. An opportunity can become a career. A community can become a global network. And one person's willingness to serve can create a bridge that continues to benefit others long after they have moved on.

Against the backdrop of the theme "Creating Global Connections," my keynote will explore the deeper meaning of connection: connecting generations, disciplines, institutions, countries, knowledge and opportunity. It will challenge women in statistics and data science to look beyond individual achievement and consider the bridges they are building for those who will come after them.

Ultimately, my keynote will invite participants to reconsider what it means to succeed. Our degrees, positions, publications, and accomplishments may tell the story of where we have been, but the bridges we build for others tell the story of what we leave behind. The question is not only, "How far can I go?" but also, "Who can go further because I chose to build a bridge?"

**ORGANIZER:** IDWSDS Program Committee**CHAIR:** Vanda Lourenço, NOVA School of Science and Technology



Carine Malor

Ghana Statistical Association

Carine is a statistician, project management professional, strategic leader, and mentor with a strong commitment to advancing the statistics profession and developing the next generation of women leaders.

She currently serves as the General Secretary of the Ghana Statistical Association and Country Coordinator for the West Africa Young Statisticians Association. She previously served as General Secretary of Women in Statistics, Ghana, from 2020 to 2024.

Carine holds a PhD in Statistics and has built her professional journey across statistics, data management, project management, leadership, professional development, and community service. Beyond her formal career, volunteering and service have played a significant role in shaping her leadership journey, creating opportunities to connect with people, ideas, and communities across different professional and geographical contexts.

She is also the founder of SheLeads Academy, an NGO through which she coaches and mentors young people, especially women to develop their career goals and leadership skills. Her work reflects her belief that leadership extends beyond personal achievement to creating opportunities, opening doors, and building pathways for others to succeed.

Carine is passionate about service, leadership, professional development, mentoring, and meaningful connections. Her experiences have shaped her belief that some of the most transformative opportunities in a career can emerge from a willingness to serve and contribute beyond one's formal role.

She brings this perspective to her keynote, "Building Bridges That Outlive Us: Creating Global Connections Through Service, Leadership and Purpose," where she reflects on her journey and the role of service, leadership, and meaningful connections in creating lasting impact.

For Carine, a successful career is measured not only by how far we go, but also by how far others can go because we chose to build a bridge.



Session Info

K6

KEYNOTE SESSION

October 6th, 23:00 - 24:00 UTC

Connecting Cure Models and Women's Health: Survival Analysis for Uterine Cancer Data

USE SCHED TO ACCESS THE ZOOM ROOM

Survival analysis plays a fundamental role in health research, especially when the objective is to understand not only the time until an event occurs, but also the possibility that a proportion of individuals may never experience that event. This situation is particularly relevant in cancer studies, where long-term survivors may be considered cured or no longer susceptible to recurrence or death related to the disease.

In this talk, I will present statistical modeling approaches based on cure models, with emphasis on their application to uterine cancer data. The discussion is motivated by my work on regression modeling for cure factors using the reparametrized defective generalized Gompertz distribution. This framework allows the estimation of both survival behavior and cure proportions, providing a flexible tool for analyzing time-to-event data in the presence of a cured fraction.

The presentation will highlight how cure models can contribute to women's health research by identifying factors associated with long-term survival and by offering interpretable measures for clinical and epidemiological studies. By connecting statistical theory, survival analysis, and real health data, this work illustrates how statistical science can support evidence-based understanding of cancer outcomes and strengthen global connections between data science and public health.

ORGANIZER: IDWSDS Program Committee**CHAIR:** Jessica Stockdale, Simon Fraser University



Vera Tomazella

Universidade Federal de São Carlos

Professor Vera Lucia Damasceno Tomazella is a Full Professor and Senior Professor at the Federal University of São Carlos (UFSCar), Brazil, affiliated with the Department of Statistics. She holds a degree in Mathematics from the Federal University of Maranhão, a Master's degree in Computer Science and Computational Mathematics from the University of São Paulo, and a Ph.D. in Computer Science and Computational Mathematics from the University of São Paulo. She completed two international postdoctoral fellowships: one at the University of Valencia, Spain, and another at the University of Manchester, England. Her research focuses on survival analysis, reliability, cure models, risk models, Bayesian inference, diagnostic analysis, and nonlinear models, with applications in health sciences and cancer data. Professor Tomazella is an elected member of the International Statistical Institute (ISI) and has actively contributed to the statistical community through research, teaching, mentoring, and scientific leadership. She has also held important professional roles, including leadership positions in Brazilian statistical and biometric societies. Her work has contributed to the development and application of statistical models for complex survival data, especially in biomedical and public health contexts.

Concurrent Sessions



Session Info

S101

OPENING SESSION

October 6th, 00:00 - 00:30 UTC

Opening Session

USE SCHED TO ACCESS THE ZOOM ROOM

This session will welcome attendees, discuss the history of the IDWSDS Conference, and you'll hear from the representatives from the Caucus for Women in Statistics & Data Science (CWS), International Statistics Institute (ISI), Korean International Statistical Society (KISS), and the International Turkish Statistical Association (ITSA). They will share how their organizations help their members create global connections. We will also share technical information about how to access key features of the Sched App and our Zoom rooms.

Speakers include:

Dr. Suhwon Lee (CWS)

Dr. Fabrizio Ruggeri (ISI)

Dr. Mi-Ok Kim (KISS)

Dr. Sinem Demiri (ITSA)

ORGANIZER: IDWSDS Program Committee

CHAIR: Samantha Seals, University of West Florida



Session Info

S102

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 00:30 - 1:00 UTC

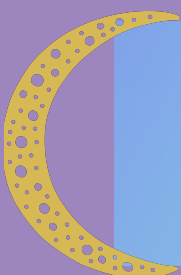
Gini-Weighted Risk Control in Adaptive Generative Augmentation

USE SCHED TO ACCESS THE ZOOM ROOM

Generative data augmentation is widely used to address class imbalance by enriching minority classes with synthetic samples. Existing approaches typically employ a fixed augmentation strength across all classes, ignoring differences in class imbalance, data structure, and generative quality. We propose an adaptive augmentation framework that determines class-specific augmentation strengths using both class proportions and structural information measured through Gini correlation. This strategy allocates more synthetic data to underrepresented and weakly structured classes while limiting augmentation for well-represented classes.

We develop a theoretical framework showing that the resulting excess risk is controlled by a weighted combination of class-conditional Wasserstein discrepancies and Gini-based structural factors. We further establish consistency results and demonstrate that adaptive augmentation provides tighter control of risk distortion than fixed augmentation schemes.

Experiments on imbalanced classification datasets show consistent improvements in minority-class recall and macro-F1 performance, while empirical results closely match theoretical predictions. The proposed framework provides a principled, structure-aware foundation for generative data augmentation.



ORGANIZER: Chathurika Abeykoon, Rhodes College

CHAIR: Margaret Banker, Northwestern Feinberg School of Medicine



Speaker Bios

S102



DR. CHATHURIKA ABEYKOON

Rhodes College

Chathurika Abeykoon is an Assistant Professor of Statistics at Rhodes College, Memphis, TN. Dr. Abeykoon received her Ph.D. in Mathematics with a concentration in Statistics from the University of Mississippi in 2023. Her research lies at the intersection of Mathematics, Statistics, and Computer Science, focusing on the theoretical foundations of machine learning and deep learning. Recently, she has been exploring the polynomial representation of shallow neural networks, double descent behavior in neural networks for classification, and Generative Augmentation for class imbalance.



Session Info

S103

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 2:00 - 3:00 UTC

Decoding Health: AI and Big Data at the Frontier of Precision Medicine

USE SCHED TO ACCESS THE ZOOM ROOM

This session explores how artificial intelligence and data science are driving the next generation of precision medicine, drawing on large-scale biobanks, electronic health records (EHRs), and genomic data from both academia and industry. Dr. Marie Loh (Nanyang Technological University / Genome Institute of Singapore) will discuss translational opportunities from HELIOS-SG100K, a multi-ethnic Asian population cohort with rich phenotypic and molecular data, with examples in cardiometabolic disease and atopic dermatitis. Ms. Chih-Ting Yang (Vanderbilt University) will use the All of Us Research Program to show how biobank data can be translated into longitudinal disease staging, illustrated through a study of cardiovascular-kidney-metabolic syndrome. Dr. Sarah Lotspeich (Wake Forest University) will present an LLM-enhanced, ICD-10-based algorithm for recovering missing EHR data, demonstrating accuracy comparable to expert chart review at scale. Rounding out the session, Dr. Christine Hsiung will bring an industry perspective from the genetics sector, sharing applied research insights from Taiwan Biobank genomic data and translational pipelines. Together, these talks span academic and industry approaches, genomic profiling, longitudinal disease modeling, and AI-driven data quality methods, illustrating how data science is reshaping precision medicine across diverse populations and settings.

ORGANIZER: Yi-Ju Lee, Institute of Statistical Science, Academia Sinica

CHAIR: Yi-Ju Lee, Institute of Statistical Science, Academia Sinica



Speaker Bios

S103



PROF. SARAH LOTSPEICH
Wake Forest University

<https://www.sarahlotspeich.com/>

Sarah Lotspeich is an Assistant Professor in Statistical Sciences at Wake Forest University. She co-leads the Spatial and Environmental Statistics in Health (SESH) Lab at Wake Forest and the Missing and INcomplete Data (MIND) Lab at UNC Chapel Hill, and is enthusiastic about mentoring student research. She co-organizes Florence Nightingale Day at Wake Forest annually, engaging local students in statistics and data science, and holds elected positions in the Caucus for Women in Statistics and Data Science. Sarah completed a postdoctoral fellowship in Biostatistics at UNC Chapel Hill and earned her Ph.D. in Biostatistics from Vanderbilt University. Her research tackles challenges in analyzing error-prone observational data, focusing on international HIV cohorts, electronic health records, and health disparities, and she develops methods for statistical modeling with censored covariates. When not writing code, she enjoys cross-stitching, traveling, and rewatching favorite TV shows.



PROF. MARIE LOH
Nanyang Technological University

<https://dr.ntu.edu.sg/entities/person/Marie-Loh-Chiew-Shia>

Marie Loh is an Assistant Professor in the Lee Kong Chian School of Medicine at Nanyang Technological University, Senior Research Scientist at the Genome Institute of Singapore and Honorary Senior Lecturer at Imperial College London. Asst Prof Loh is a molecular epidemiologist with a long-standing interest in the role of genetics, epigenetics and other omics underlying complex diseases, with a specific focus on DNA methylation and Asian populations. Her current research focuses on combining population health with molecular phenotyping approaches to advance knowledge and understanding in non-communicable diseases, including cardiometabolic diseases and atopic dermatitis. She has published more than 100 papers to date, including high-impact journals such as Nature and Nature Genetics.



MS. CHIH-TING YANG
Vanderbilt University

<https://www.vumc.org/biostatistics/person/chih-ting-teresa-yang>

Chih-Ting Yang is a PhD candidate in Biostatistics at Vanderbilt University. Her research focuses on developing and applying statistical methods to complex biomedical data, including microbiome and single-cell omics data, electronic health records, and biobank-linked clinical data.



DR. CHIANI HSIUNG
Genetics Generation Advancement Corp.(GGA Corp.)

Dr. Chia-Ni Hsiung (Christine Hsiung) is a Taiwan-based researcher in genomic epidemiology and precision medicine. She holds an MPH from the National Defense Medical Center and a Ph.D. in Precision Medicine from National Tsing Hua University (2026). She currently serves as Senior Deputy Project Manager at GGA Corp..

Her research applies GWAS, multi-trait analysis, and Mendelian randomization to ophthalmic comorbidities, chronic kidney disease, breast cancer susceptibility, and metabolic traits — including a causal link between depression and cataract risk, an East Asian-specific polygenic risk score for kidney function decline, and contributions to international breast cancer GWAS studies published in Nature Genetics.



Using large language models to enhance clinically-driven missing data recovery algorithms in electronic health records

Sarah Lotspeich, Wake Forest University

Electronic health record (EHR) data are prone to missingness and errors. We previously devised an enriched chart review protocol using a "roadmap" of auxiliary diagnoses to recover missing values. However, chart reviews are expensive and time-intensive, limiting how many patients' data can be reviewed. We now investigate the accuracy and scalability of a roadmap-driven algorithm, based on ICD-10 codes, designed to mimic expert chart reviews and recover missing values. In addition to the clinicians' original roadmap, we consider new versions iteratively refined using large language models (LLMs) with clinical expertise to expand the list of auxiliary diagnoses. Using chart reviews for 100 patients from an extensive EHR, we examine algorithm performance. Across these patients, there were 413 missing values. Expert chart reviews recovered 49 (12%), while LLM-enhanced roadmaps recommended nearly twice as many (83-89, 20%-22%). The final algorithm, using clinician-approved LLM additions, offered a balance (73, 18%), expanding the roadmap with LLM suggestions only when clinically relevant. Applied to a larger study of 1000 patients from the same EHR, the per-patient median number of non-missing values increased from 6 to 7. Clinically-driven algorithms enhanced by LLMs can recover missing EHR data with accuracy comparable to chart reviews and can feasibly scale to large samples. Extending these methods to other data quality dimensions is a promising future direction.

Translating Research from a Multi-Ethnic Asian Population Cohort

Marie Loh, Lee Kong Chian School of Medicine, Nanyang Technological University

In this short talk, Assistant Professor Marie Loh will be sharing with us the opportunities and challenges that lies around the use of the rich phenotypic and molecular data available within the HELIOS-SG100K study, a multi-ethnic Asian population cohort. She will highlight the potential for translational opportunities, illustrating with several examples across different disease domains.

Leveraging the All of Us Research Program for Longitudinal Disease Staging to Support Precision Medicine

Chih-Ting Yang, Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

Large-scale biobank resources provide new opportunities for precision medicine by integrating clinical, survey, genomic, and other biomedical data from diverse populations. In this talk, I will use the All of Us Research Program as an example to discuss how data science methods can help translate biobank data into longitudinal disease information.

As an example application, I will present a study of cardiovascular-kidney-metabolic syndrome, where we use electronic health record data to move beyond static baseline disease classification and characterize disease stages over time.

This example highlights several key challenges in biobank-based research, including longitudinal staging, irregular clinical measurements, informative missingness, and interpretation of disease trajectories. More broadly, the talk will discuss how biobank data can support disease progression research, risk stratification, and precision medicine.

Genetic Insights into Ocular Disorders and Comorbid Diseases: A Genome-Wide Association and Polygenic Analysis

Chiani Hsiung, Genetics Generation Advancement Corp.(GGA Corp.)

Eye diseases, including cataracts, glaucoma, diabetes retinopathy, and age-related macular degeneration, are major global health challenges and leading causes of blindness. This study leveraged genome-wide association studies (GWASs) involving over 100,000 individuals, integrating data from the Taiwan Biobank and National Health Insurance Research Database, to identify genetic loci associated with disease onset. Our findings suggest that these conditions are influenced by multifactorial etiologies, as pleiotropic loci including rs10811660, rs4710941, rs2283228, and rs7646518 were identified, linking ocular diseases to metabolic conditions. Notably, a strong genetic correlation was observed between cataract and depression. Mendelian randomization analysis further demonstrated a causal effect of depression on cataract risk, implicating shared biological pathways, particularly oxytocin signaling, in disease pathophysiology. This finding revealed a functional genetic variant near the OXTR gene, highlighting its potential as a causal candidate for genetic diagnosis in precision health. By bridging the gap between genetic discovery and clinical application, this research offers critical insights into shared genetic mechanisms across diverse health domains, paving the way for innovative diagnostic and therapeutic strategies.



Session Info

S104

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED SESSION

October 6th, 3:00 - 4:00 UTC

Breaking Barriers

USE SCHED TO ACCESS THE ZOOM ROOM

An Invited Session Proposal from Statistics Society of Australia (SSA) Women in Statistics Network Supported by the The International Statistical Institute (ISI) Committee on Women in Statistics

The progression of the talks:

curiosity → identity → contribution → legacy
aspiration → exploration → contribution → impact

Speakers:

1. Jayamini Liyanage, j.liyanage@latrobe.edu.au, La Trobe University, Australia, (PhD Student)
2. Alysha De Livera, PhD, AStat, A.DeLivera@latrobe.edu.au, Latrobe University, Australia, (Mid Career)
3. Ayse Aysin Bilgin, PhD, ayse.bilgin@mq.edu.au, Macquarie University, Australia Professor (Late Career)

ORGANIZER: Alysha De Livera, La Trobe University

CHAIR: Ayse Aysin Bilgin, Macquarie University, Australia



Speaker Bios

S104



MRS. JAYAMINI LIYANAGE
La Trobe University

<https://scholars.latrobe.edu.au/jliyanage/about>

Jayamini Liyanage is a final-year Biostatistics PhD student at La Trobe University, Melbourne, where her research focuses on developing multivariate meta-analysis methods for analysing high-dimensional biological data. She has published statistical methods and developed an R package to support evidence synthesis for high-dimensional data and has presented her work at multiple conferences and events. Alongside her research, Jayamini actively engages with the School of Computing, Engineering and Mathematical Sciences community at La Trobe University and works as a casual academic. She is passionate about making statistics more accessible and encouraging greater participation and belonging in STEM fields.



DR. ALYSHA DE LIVERA
La Trobe University

<https://scholars.latrobe.edu.au/a2delivera>

Dr Alysha De Livera is Co-Chair of the newly established Women in Statistics and Data Science Special Interest Group of the Statistical Society of Australia, and previously served as Co-Chair of its Biostatistics and Bioinformatics Committee. She is an academic in Statistics in the Department of Mathematics and Statistics at La Trobe University, Melbourne, where she is Director of the Statistical Consultancy Platform and Partnership Coordinator for Mathematical and Physical Sciences. Her work spans methodological and collaborative statistical research, and she co-leads the Dimension Reduction and Meta-Analysis Research Group at La Trobe University. Alysha is actively engaged in teaching, mentoring, consulting, and student supervision, and is a strong advocate for women in STEM.



PROF. AYSE AYSIN BILGIN
Macquarie University

<https://researchers.mq.edu.au/en/persons/ayse-bilgin/>

Honorary Prof Ayse Aysin Bilgin is a Vice President of International Statistical Institute (2025-2029), co-chair of Statistics Education Section of Statistical Society of Australia and was the President of the International Association for Statistical Education (IASE). She has held a range of academic leadership roles, contributing to curriculum design, program development, and institutional strategy in teaching and learning at Macquarie University in Australia. Bilgin is recognised for her contributions to the scholarship of teaching and learning in statistics, with publications and invited talks that influence both academic practice and policy. Through her leadership and research, she continues to advocate for inclusive, evidence-based approaches to education that equip students with the skills needed in an increasingly data-centric world.





Breaking barriers to embrace the unknown

Jayamini Liyanage, La Trobe University, Australia

Beginning a PhD often means stepping into the unknown: unfamiliar methods, complex data, new research expectations, and an uncertain path ahead. For me, this journey also involved moving countries to pursue research in a new academic environment, which brought its own challenges of adaptation, confidence-building, and finding a sense of belonging. In this talk, I reflect on my journey as a PhD student working on multivariate meta-analysis methods for high-dimensional biological data. I share how curiosity has helped me navigate uncertainty, from learning new statistical ideas to developing tools that support researchers in evidence synthesis. Along the way, I have come to see the unknown not simply as a barrier, but as a space where growth begins. This talk is about embracing uncertainty as part of the journey. It is about how curiosity can turn uncertainty into possibility, how unfamiliar environments can become places of growth, and how early-career statisticians can begin contributing meaningfully while still learning, growing, and discovering the path ahead.

Breaking barriers to leave a trail

Alysha De Livera, La Trobe University, Australia

Every statistical project leaves a trail: of questions asked, methods developed, collaborations built, and stories told. Yet we don't often stop to think about what it means to leave a meaningful trail for others to follow. In this talk, I reflect on the barriers and opportunities that have shaped my journey as an academic statistician and a woman of colour, working across research, consulting, and teaching, and how these experiences continue to shape the way I move through and contribute to the field. I also consider what it means to leave a trail: not just doing the work, but making the path clearer, more visible, and more supportive for those who come next!

Breaking barriers to reach goals!

Ayse Aysin Bilgin, Macquarie University, Australia

It is very exciting to be at the beginning of one's career in Statistics. There are a lot of possibilities, and a lot of opportunities to excel. Finding the dream job, setting the right goals, adjusting the goals for changed circumstances and achieving them is easy with good mentors and right attitude. Interconnectedness, collaboration and working together helps a lot. Would you like to hear from an academic in Statistics who were able to find ways to get help whenever she needed? Would you like to learn from my experiences so that your journey towards the success is much faster? If so, come along to learn and benefit from my experiences of how to achieve your best despite the barriers on your way to top!



Session Info

S105

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 4:00 - 5:00 UTC

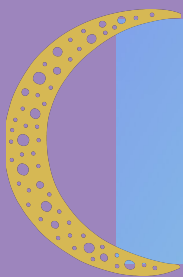
Towards Inclusive Medical AI: Mitigating Data Bias and Fostering Equity

USE SCHED TO ACCESS THE ZOOM ROOM

As artificial intelligence becomes deeply integrated into healthcare, addressing inherent biases and ensuring equitable outcomes is paramount. This session explores multidisciplinary strategies for developing inclusive and fair medical AI systems. The panel brings together four experts to discuss critical challenges and actionable solutions. Dr. Heajin Kim (Chair) will highlight the importance of reflecting sex and gender characteristics in medical AI to foster equity. Researcher Jeongyeon Kim will provide a critical analysis of data bias issues within Korea's AI Hub. Dr. Yuseong Chu will review current research trends and methodologies for achieving algorithmic fairness in healthcare. Finally, Dr. Wonjung Park will introduce innovative approaches utilizing synthetic data generation to address dataset imbalances. Together, this session offers a comprehensive roadmap towards trustworthy and inclusive medical AI.

ORGANIZER: Heajin Kim, Center for Gendered Innovations in Science and Technology Research (GISTeR)

CHAIR: Heajin Kim, Center for Gendered Innovations in Science and Technology Research (GISTeR)



Speaker Bios

S105



DR. HEAJIN KIM

Korea Foundation for Science Culture and Innovation

Dr. Heajin Kim is a Senior Researcher at the Center for Gendered Innovations in Science and Technology Research (GISTeR). She earned her Ph.D. in Bio-convergence and Neuroscience from the University of Science and Technology (UST) and the Korea Institute of Science and Technology (KIST). Her current research focuses on mitigating bias and fostering equity in emerging healthcare technologies, as well as analyzing sex differences in neuroscience.



MISS JEONGYEON KIM

Korea Foundation for Research Culture and Innovation

Jeongyeon Kim is a Researcher at the Center for Gendered Innovations in Science and Technology Research (GISTeR). She earned her bachelor's degree with a double major in Statistics and Information Technology. Her current work focuses on analyzing bias and representativeness in healthcare datasets, with particular attention to sex and gender disparities in medical AI data.



DR. YUSEUNG CHU

Konyang University

Dr. Yuseung Chu received his B.S. and Ph.D. degrees in Biomedical Engineering from Yonsei University. He is currently a Postdoctoral Researcher in the Department of Precision Medicine at Yonsei University Wonju College of Medicine. His research focuses on medical artificial intelligence, particularly in medical image analysis, dermatology AI, multimodal learning, and generative AI. He is especially interested in developing fair and robust AI models that can reduce performance disparities across demographic and clinical subgroups. His recent work explores bias, fairness, and reliability in skin disease classification and skin lesion image generation using diverse dermatological datasets.



DR. WONJUNG PARK

KAIST

Dr. Wonjung Park is a Postdoctoral Researcher in the Computer Graphics and Visualization Lab at KAIST. She received her Ph.D. in Computer Science from KAIST in 2026. Her research specializes in NeuroImaging and Generative AI, with a particular focus on the biological trustworthiness of medical image generative models and object-oriented synthetic data generation.





Integrating Sex and Gender Characteristics into Medical AI for Enhanced Equity

Heajin Kim, Senior researcher

This presentation highlights the critical need to reflect sex and gender differences in the development of medical artificial intelligence. By exploring the limitations of gender-blind AI models, this talk will propose actionable strategies for integrating sex and gender characteristics to mitigate bias and foster equity in healthcare outcomes.

Exploring Data Bias: An Overview of Korea's AI Hub Healthcare Datasets

Jeongyeon Kim, Researcher

High-quality data is the foundation of fair AI. This talk presents an overview of publicly available healthcare datasets in Korea's AI Hub, focusing on demographic representation and potential sources of data bias. By examining dataset characteristics and distribution patterns, the presentation shares key observations and raises considerations for the development of equitable medical AI.

Methodological Advances in Fair Medical AI: A Review of Current Research Trends

Yusung Chu, Yonsei University, Postdoctoral Researcher

This presentation reviews and summarizes recent studies on algorithmic fairness in medical AI. It focuses on how bias and data imbalance are identified and measured, and how performance differences across patient subgroups are evaluated. The talk will also introduce representative approaches to reduce bias, such as subgroup-aware evaluation, data rebalancing, and algorithmic debiasing, providing an overview of current efforts toward more inclusive medical AI.

Addressing Dataset Imbalances through Object-Oriented Synthetic Data Generation

Wonjung Park, Postdoctoral Researcher, School of Computing, KAIST

Data scarcity and imbalance are significant hurdles in training inclusive medical AI. This talk explores innovative techniques utilizing synthetic data generation to address these imbalances. Drawing on recent developments in generative AI, the presentation will demonstrate how object-oriented synthetic data can enhance the diversity and biological trustworthiness of medical datasets.



Session Info

S106

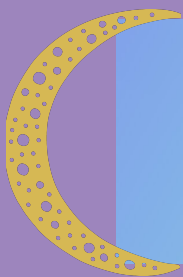
CONTRIBUTED SESSION

October 6th, 5:00 - 6:00 UTC

Statistical Methods and Applications 1

USE SCHED TO ACCESS THE ZOOM ROOM

Refining a Two-Stage Pipeline for Gene-by-Environment Interaction Discovery:
Random-Forest Screening Followed by SOIL-Based Prioritization | Mapping Inequality:
Women Statisticians and the Settlement House Movement | Precision by Design: A
Transferable Framework for Sampling Optimisation in Hierarchical
Capture-Recapture | How Biostatistics Research Experience Enriches the Pre-Medical
Journey

ORGANIZER: IDWSDS Program Committee**CHAIR:** Sachin Sachdeva, University of Hyderabad

Speaker Bios

S106



MRS. KAKON DATTA
University of Kentucky

Kakon Datta is a Ph.D. student in the Department of Statistics at the University of Kentucky. She completed her master's degree at Miami University, Ohio, USA in 2024. Her research interests include high-dimensional data analysis, variable selection, sufficient variable selection, and statistical learning methods. She is particularly interested in developing efficient methods for detecting important variables and interactions in complex datasets.

Alongside her research, Kakon serves as a graduate assistant and has experience teaching undergraduate statistics courses. She is passionate about applying modern statistical techniques to real-world problems and enjoys learning new methodologies in computational and theoretical statistics. Outside academics, she enjoys reading books and exploring new ideas across different fields.



DR. SU NA CHIN
Universiti Malaysia Sabah

Dr. Chin Su Na is a statistician and Lecturer in Statistics at the Faculty of Science and Technology, Universiti Malaysia Sabah, Malaysia. Her research focuses on developing statistical methodology for efficient sampling design and population estimation. She received her PhD in Mathematical Sciences from the University of Southampton, United Kingdom, where she developed a unified framework for optimising sampling effort in capture-recapture studies. Her research combines mathematical statistics, optimisation, and computational methods to address challenges in statistical study design under resource constraints. While motivated by capture-recapture applications, her work has broader relevance to survey methodology, ecological statistics, public health, and other fields that depend on efficient data collection and reliable population estimation.



MS. TISHA PRASAD
STEM4Change

Tisha Prasad is the co-founder and co-president of STEM4Change, a nonprofit organization dedicated to expanding access to STEM education by teaching students foundational concepts like coding. She also serves as Head of Web Development at VolunteerConnect, where she leads the development of technology that connects volunteers with nonprofit organizations and service opportunities. As a Women in Data Science ambassador, she is leading the development of Lift Our Voices, a project focused on closing the gender data gap to improve the accuracy and inclusivity of automatic speech recognition across diverse vocal patterns. Her interests lie at the intersection of technology, education, and social impact, with a focus on developing equitable technologies that empower communities and broaden opportunities for learning.

<https://www.linkedin.com/in/kevinzoufl>



MR. KEVIN ZOU
University of Central Florida

Kevin Zou is an undergraduate student majoring in Health Sciences (B.S.) on the Pre-Clinical Track at the University of Central Florida (Class of 2029). He is an Undergraduate Researcher in Dr. Julia Soulakova's biostatistics research team, where his work explores public health disparities across a range of topics, including sleep duration, unpaid caregiving, and psychological well-being among Florida college students. Kevin also works as a Certified Nursing Assistant in the Intensive Care Unit at Halifax Health. Drawing on both experiences, he is passionate about how biostatistics research and clinical work shape students' preparation for a future career in medicine.



Refining a Two-Stage Pipeline for Gene-by-Environment Interaction Discovery: Random-Forest Screening Followed by SOIL-Based Prioritization

Kakon Datta, University of Kentucky

Detecting gene-environment (G×E) interactions is challenging because interaction effects are typically weak, genetic data are high-dimensional, and environmental effects may be complex and heterogeneous. We propose a two-stage framework that combines nonparametric screening with sparse, interpretable modeling for high-dimensional G×E analysis. First, SNPs are aggregated within gene-level sets, and a random forest is used to screen and rank candidate gene-environment interactions. Second, the retained candidates are analyzed with a sparse regression model including both main effects and interaction terms, producing effect estimates and an interpretable interaction structure. To improve stability, we incorporate uncertainty-aware interaction importance based on model-averaged inclusion across candidate sparse models. The method is motivated by the sparsity of true interaction signals and the need to reduce the large search space before formal inference. Preliminary investigations indicate stable performance under diverse G×E settings, and further results from simulation studies and a depression-related application will be presented.

Mapping Inequality: Women Statisticians and the Settlement House Movement

Tisha Prasad, STEM4Change

This talk will examine the connection between the Settlement House movement and the development of social statistics, focusing on the women who used data collection and visualization to drive labor and welfare reform in late nineteenth-century Chicago. At Hull House, reformer Florence Kelley and her colleagues (including Jane Addams, Julia Lathrop, and Agnes Sinclair Holbrook) conducted door-to-door surveys of one of the city's most overcrowded immigrant districts, mapping wages and nationalities block by block. The resulting 1895 publication, Hull-House Maps and Papers, became a landmark in statistical graphics and a model for the modern social survey. Far from a purely academic exercise, this data collection was deployed directly as evidence: Kelley's findings on child labor and sweatshop conditions were presented to Congress and the Illinois legislature, helping secure passage of the state's first factory inspection laws. This talk traces how Hull House women built one of the earliest bridges between statistical methodology and social policy, and considers their legacy for the role of data in advocacy today.

Precision by Design: A Transferable Framework for Sampling Optimisation in Hierarchical Capture-Recapture

Su Na Chin, Universiti Malaysia Sabah

Capture-recapture methods are widely used to estimate population size when complete enumeration is impractical. Yet sampling effort across capture occasions is often allocated by convention rather than statistical efficiency. This study develops a hierarchical Schnabel census in which each primary occasion

contains several sub-occasions, such as trapping days or repeated deployments. Under a fixed total effort, the aim is to determine the allocation that minimises the variance of the population estimate. When per-sub-occasion capture probability is constant, convexity arguments and Jensen's inequality show that equal allocation is optimal. Simulations confirm that it generally yields lower empirical and asymptotic variance than skewed or random allocations. The same result holds when capture probabilities follow a common symmetric distribution. However, when capture probabilities decline over time, equal allocation is no longer optimal. Monte Carlo optimisation indicates that more effort should be assigned to later occasions to offset reduced detectability and improve precision. The findings show that efficient design depends not only on total effort, but also on how detectability changes over time. The framework provides practical guidance for designing more efficient capture-recapture studies across ecology, fisheries, conservation, and related fields.

How Biostatistics Research Experience Enriches the Pre-Medical Journey

Kevin Zou, University of Central Florida

Pursuing a career in medicine, public health, or another discipline often requires students to simultaneously build a competitive graduate school application while developing the knowledge, skills, and attributes needed to become future clinicians, scientists, or other professionals. At first sight, biostatistics research may appear distant or unrelated to a future career, but many of the skills it builds are directly transferable. Participation in biostatistics research offers a unique opportunity to cultivate technical, analytical, and writing skills. As an undergraduate student, I will reflect on my experience of working within a biostatistics research team on a project investigating disparities in sleep duration among Florida college students. In this talk, I will share my perspective on how to find the right biostatistics lab and how my research experience – from learning to code with SAS® 9.4 to presenting the work at a conference – supports my pre-medical development and goes beyond the resume. I will conclude with an overview of the broader value of biostatistics research for pre-medical and non-statistics students in general, emphasizing that research extends well beyond merely strengthening applications.



Session Info

S107

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

PLENARY TALK

October 6th, 7:00 - 7:30 UTC

Creating Global Connections Through Early Statistical Leadership in Rare Disease Trials

USE SCHED TO ACCESS THE ZOOM ROOM

In complex clinical research, collaboration begins well before database lock or final analysis. It begins when statisticians are involved early enough in protocol development to shape what is measurable, sustainable, and meaningful for patients. This presentation uses a rare pediatric dermatology trial as a case study to show how early statistical input can prevent avoidable missing data before the first participant is enrolled.

The original design was scientifically ambitious but operationally burdensome, with frequent in-clinic visits, narrow assessment windows, and substantial caregiver burden. Statistically, this created a foreseeable risk of informative missingness, loss to follow-up, and reduced interpretability in a small and vulnerable population. Early statistical leadership helped reframe the design around feasibility as well as rigor.

The talk will illustrate how collaboration among statisticians, clinicians, programmers, operations teams, and digital partners can reduce this risk. Examples include identifying burden-sensitive endpoints, anticipating missing-data pathways, and considering remote ePRO, image capture, or AI-supported skin assessment to replace selected in-person visits without compromising data quality.

The broader message is that statistics is not only an analysis discipline; it is also a design and connecting discipline.

ORGANIZER: Shrutimita Pokhariyal, PHASTAR**CHAIR:** Pamela Solano, Regensburg University

Speaker Bios

S107



MS. SHRUTIMITA POKHARIYAL
PHASTAR

Shrutimita Pokhariyal is an experienced biostatistics professional with over 14 years in clinical research, with expertise spanning multiple phases of clinical trials. Her work includes statistical strategy, study design, protocol and SAP development, and collaboration across cross-functional teams to support high-quality research delivery. She has been an invited speaker at both national and international forums, where she has shared insights on statistics, innovation, and clinical development. She is also committed to mentoring aspiring professionals and advancing women's leadership in STEM.



Session Info

S108

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 7:30 - 8:00 UTC

Explainable AI in Health Technology

USE SCHED TO ACCESS THE ZOOM ROOM

The rapid deployment of machine learning in healthcare has exposed a fundamental tension between model performance and clinical utility. This talk addresses the urgent need for Explainability (XAI), moving beyond the "black box" paradigm to ensure that AI systems are not only accurate but also transparent, accountable, and trustworthy. In the high-stakes environment of medical diagnostics, understanding an algorithm's pathway is essential for verifying clinical reliability and meeting the rigorous standards demanded by both clinicians and regulators.

A central theme of this talk is the transition from surface-level performance metrics to deep reproducibility and validation. We will examine how traditional accuracy scores can sometimes be deceptive and how to combat this, compare XAI with true interpretability, and tackle the challenge of the inheritance of bias. This session outlines proactive mitigation strategies, including rigorous data quality checking, subgroup analysis, and continuous fairness assessments throughout the model lifecycle.

Finally, the talk will navigate the evolving regulatory landscape, specifically the EU AI Act.

ORGANIZER: Autumn Johnson, University of Galway**CHAIR:** Pamela Solano, Regensburg University

Speaker Bios

S108



DR. AUTUMN JOHNSON

University of Galway

My name is Autumn Johnson, and I'm a postdoctoral researcher in Statistics at the University of Galway. My postdoctoral work has focused on health technology advancements using complex statistical methods, machine learning, and AI. My PhD is from University College Cork, where I graduated in 2025. My research focused on applying machine learning to incorporate genomic data into survival analysis for prostate cancer prognosis.



Session Info

S109

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 8:00 - 9:00 UTC

Advances in Statistical Methods for Complex and High-Dimensional Data

USE SCHED TO ACCESS THE ZOOM ROOM

This session brings together recent advances in statistical methodology for analyzing complex, high-dimensional, and heterogeneous data arising in modern biomedical and genetic research. The presentations cover a broad range of topics, including variable screening for mixed-type and high-dimensional data, integration of mechanistic and statistical models for infectious disease prediction, semiparametric approaches for meta-analysis, and set-based association testing for functional genetic responses. Together, these talks highlight emerging statistical frameworks that improve flexibility, scalability, and interpretability in the analysis of complex data structures.

ORGANIZER: Seo Young Park, Korea National Open University**CHAIR:** Seo Young Park, Korea National Open University

Speaker Bios

S109



PROF. HAEUN MOON
Seoul National University

Haeun Moon is an assistant professor in the Department of Transdisciplinary Innovations and the Department of Statistics at the Seoul National University, South Korea. Before joining SNU, she was a postdoctoral researcher in the Department of Statistics and Data Science at Carnegie Mellon University. She received Ph.D in Statistics from the University of Pittsburgh. Her research focuses on developing association measures for modern data, which are often characterized by nonlinear relationships, heterogeneous data types, missingness, and non-i.i.d. structures. She is also interested in leveraging black-box prediction models for powerful yet valid statistical inference. Her applied work centers on association studies in genomics.



PROF. SAEBOM JEON
Sungshin Women's University

Saebom Jeon is an associate professor in the School of Mathematics, Statistics and Data Science at Sungshin Women's University and a research member of the Biomedical Mathematics Group at the Institute for Basic Science (IBS), Korea. She received her Ph.D. in Statistics from Korea University. Her research focuses on statistical modeling for complex and high-dimensional data, including latent variable models, longitudinal and trajectory analysis, and multiscale integration of mechanistic and statistical approaches, with applications in public health, epidemiology, and biomedical data science.



PROF. EUNJEE LEE
Chungnam National University

Eunjee Lee is an Associate Professor in the Department of Information and Statistics at Chungnam National University. Her research focuses on functional data analysis, biomedical imaging, brain network analysis, and Bayesian methodology. She has developed statistical models for complex biomedical data, including hippocampal shape imaging and functional brain connectivity, with applications to Alzheimer's disease and neuroimaging studies.



PROF. SUNYOUNG SHIN
Pohang University of Science and Technology

Sunyoung Shin is an Associate Professor in Department of Mathematics at Pohang University of Science and Technology. Her research lies at the intersection of statistical learning and reinforcement learning, with a focus on developing scalable methods for high-dimensional data, particularly in genomics and bioinformatics. Her work aims to uncover the biological mechanisms underlying diseases and complex traits. She received her PhD from University of North Carolina at Chapel Hill.



Frequent-voting independence screening for data of different types or different dimensions*Haeun Moon, Seoul National University*

Modern datasets often include different types of variables with complex features, making variable selection particularly challenging. For example, a measure of dependence with the response variable may not be directly comparable among predictor variables of different types and different dimensions. To address this challenge, this work proposes a frequent-voting based independent screening method for variable selection, which avoids a direct comparison of the dependence measure among different variables. Asymptotic analyses show that the proposed method selects all of the active variables with probability converging to one. We also demonstrate its great finite sample performance through numerical experiments and the application to an ADHD study.

A Multi-Scale Integration of Mechanistic and Statistical Models for RSV Prediction with ILI-Informed Calibration*Saebom Jeon, Sungshin Women's University*

Mechanistic epidemic models provide a principled framework for understanding infectious disease dynamics but often fail to capture interannual variability and short-term fluctuations. In this study, we develop a hybrid statistical-mechanistic framework to enhance a SLIRS-based model of respiratory syncytial virus (RSV) transmission.

First, we propose a season-level calibration method that incorporates influenza-like illness (ILI) information to correct systematic bias in peak amplitude and timing. Second, we model the residuals of the mechanistic model using autoregressive time-series methods, capturing temporal dependence not explained by the underlying epidemiological structure.

Using weekly RSV hospitalization data in South Korea (2015–2020), we show that ILI-informed calibration improves peak-level prediction, while residual-based autoregressive correction significantly reduces both overall and peak-period errors. Interestingly, ILI does not improve short-term predictions when included as an exogenous variable in time-series models, suggesting a separation between seasonal and short-term dynamics.

Our findings emphasize the importance of combining statistical learning techniques with mechanistic models and provide a general framework for improving epidemic prediction.

fSKAT: A Set-Based Association Test for Functional Responses in Genetic Studies*Eunjee Lee, Chungnam National University*

Set-based association testing is widely used in genetic studies, but most existing methods are developed for scalar or multivariate outcomes. In biomedical and neuroimaging studies, however, phenotypes are increasingly observed as functional data, such as signals, imaging profiles, and connectivity curves

indexed by time, space, or anatomical distance. We propose functional SKAT (fSKAT), a score-type variance-component test for assessing the association between a set of genetic variants and a functional response. fSKAT is designed to capture genetic effects that are weak at individual points but accumulate across the functional domain. The method is formulated within a function-on-scalar regression framework and constructs a kernel-based statistic from residual functions estimated under the null model, allowing inference through a mixture-of-chi-square null approximation. Simulation studies show well-calibrated type I error rates and improved power over scalar-summary-based testing approaches. We apply fSKAT to resting-state fMRI and genetic data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), using seed-based functional connectivity curves as a phenotype. The ADNI analysis examines genetic associations with early changes in brain functional connectivity before overt clinical diagnosis and illustrates the utility of fSKAT for complex functional phenotypes.

Sparse and heterogeneous meta-analysis with semiparametric models*Sunyoung Shin, Pohang University of Science and Technology (POSTECH)*

Meta-analysis is a powerful tool to synthesize findings and statistical evidence from independent studies. We consider fixed-effects semiparametric models for meta-analysis, in which different studies may have different values of finite-dimensional parameters of interest such that effect sizes are heterogeneous with the common sparse structure among studies. Nuisance parameter values may also be different. We develop a meta-analysis approach for estimating the finite-dimensional parameters and selecting features solely based on study-specific estimators via regularization employed among studies. The resulting estimators, named sparse and heterogeneous meta-analysis (SHMA) estimators, have estimation and selection consistency without access to individual-level data. Semiparametric efficiency bound is attained when the models are refitted with individual-level data after variable selection. The method is easy to implement with the alternating direction method of multipliers algorithm. This general approach has broad applications including meta-analysis of right-censored data and interval-censored data with Cox proportional hazards models. Simulation studies illustrate the superior performance of our meta-analysis approach over alternative approaches using individual-level data, such as analyses of individual studies and integrative analyses of multiple studies. Analysis of two acute myeloid leukemia survival datasets demonstrates the practical utility of our proposed approach.



Session Info

S110

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 9:00 - 10:00 UTC

Advances in Multi-Source Statistics

USE SCHED TO ACCESS THE ZOOM ROOM

This session is organized by the European Survey Research Association – Special Interest Group on multi-source statistics.

As survey research increasingly operates in a complex data ecosystem, integrating probability-based surveys with administrative records, digital traces, non-probability samples, and other non-traditional data sources has become essential for improving inference and studying phenomena that cannot be adequately measured by a single data source. In particular, new data sources can provide better data about hard-to-reach populations and minorities, helping to make these groups more visible in research and evidence-based decision-making. This session will discuss cutting-edge methodologies and innovative case studies that integrate diverse data sources in survey research, with a particular focus on their potential to improve the measurement of underrepresented populations.

ORGANIZER: Camilla Salvatore, Utrecht University

CHAIR: Camilla Salvatore, Utrecht University



Speaker Bios

S110



PROF. MONICA PRATESI
University of Pisa

https://people.unipi.it/monica_pratesi/

Monica Pratesi is Full Professor of Statistics at the University of Pisa and a member of the Scientific Council of the Italian National Research Council (CNR). She served as President of the International Association of Survey Statisticians (IASS) from 2021 to 2023 and as Director of the Department for Statistical Production at the Italian National Institute of Statistics (Istat) from 2022 to 2024, following her service on the Istat Governing Board (2016–2022). She was also President of the Italian Statistical Society (2016–2020). Her research interests include small area estimation, poverty mapping, citizen data quality, inference for hard-to-reach populations, nonresponse in telephone and web surveys, and the impact of complex survey designs on statistical modelling. She has coordinated several national and international research projects on Small Area Statistics, funded by the European Commission and the Italian Ministry of University and Research.



MX. CHAR HILGERS
German Institute for Economic Research (DIW Berlin)

https://www.diw.de/sixcms/detail.php?id=diw_01.c.890326.en

Char Hilgers is a PhD student in Sociology at the Humboldt University's Berlin Graduate School of Social Science, funded by the Socio-Economic Panel at DIW Berlin. Their research is on statistical techniques for nonresponse in survey settings: when missingness means something. From multiple imputation for handling missing not at random, to the data integration of probability and non-probability surveys. They are interested in refusal, misdirection, and obfuscation in survey settings, and ethical dilemmas around these. They hold a Master's in Industrial and Applied Mathematics and a Bachelor's in Pure Mathematics, and they have worked as a data scientist in forest monitoring, epidemiology, and supply chain forecasting.



DR. JISU KIM
Utrecht University

<https://www.uu.nl/staff/JKim>

Dr. Jisu Kim is currently an assistant professor at Utrecht University, the Netherlands in the department of Interdisciplinary Social Science. She holds a PhD in Data Science from Scuola Normale Superiore in Italy. Prior to her current position, she was a research scientist at Max Planck Institute for Demographic Research. She develops novel insights and approaches to understand migrants' trajectories, encompassing both their journeys and their engagement in digital spaces, while also examining the social, cultural, and political contexts that shape migration. By integrating computational methods with sociopolitical theory, she advances both the conceptual and empirical understanding of migration, moving beyond traditional survey-based approaches to capture real-time, large-scale patterns.



MRS. MANUELA SCHMIDT
RPTU Kaiserslautern-Landau

Manuela Schmidt is a researcher at RPTU Kaiserslautern-Landau, where she works on the German Longitudinal Environmental Study (GLEN). Her research focuses on survey methodology, geodata integration, and data quality in quantitative social research.

Towards a Quality Framework for Gender-Related Citizen Data: Challenges, Principles and Perspectives

Monica Pratesi, *University of Pisa*

Citizen-generated data (CD) is an increasingly important source of evidence for monitoring gender equality and supporting the Sustainable Development Goals (SDGs). Produced by civil society organizations, community networks, and citizens, these data can capture lived experiences and emerging issues often missed by traditional statistical systems, including gender-based violence, unpaid care work, discrimination, and the realities of marginalized populations.

However, the growing use of citizen data raises important quality challenges. Traditional quality frameworks in official statistics, based largely on probability sampling and the Total Survey Error paradigm, do not fully reflect the diversity of citizen data sources, methods, and objectives.

Gender-related citizen data is not simply sex-disaggregated data; it also reflects how power relations and inequalities shape both social realities and data production. Quality must therefore extend beyond statistical accuracy.

This paper proposes a flexible quality framework based on three dimensions: Trustworthiness, Inclusivity, and Utility. Drawing on collaborations between National Statistical Offices and Civil Society Organizations, it explores how these dimensions can support rigorous and inclusive quality principles that recognize lived experience as a valuable form of evidence for policy and advocacy.

Combining probability and nonprobability samples to estimate linear model coefficients

Char Hilgers, *German Institute for Economic Research (DIW Berlin), Socio-Economic Panel*

Nonprobability sampling is increasingly preferred for surveys due to much lower costs and ease of implementation, compared to large probability samples. However, probability samples remain important because they generally provide more accurate estimates than nonprobability samples and provide benchmark information for nonprobability surveys.

There is a growing body of research on the integration of probability and nonprobability samples to enjoy the strengths of both.

Several approaches have been developed to integrate probability and nonprobability samples, including pseudo-weighting, blended calibration, use of informative priors, and doubly robust estimators.

We present a comprehensive comparison of the pseudo-weighting, blended calibration, informative prior, and composite estimator approaches, to evaluate their efficacy in estimation of parameter values and reduction of variance in linear model coefficients, in contrast to the probability sample alone.

We apply the methods to a real-data probability and nonprobability sample measuring beliefs in conspiracy myths in Germany. We compare maximum likelihood estimates of regression coefficients and 95 percent confidence intervals for each method, with outcome variables on conspiracy beliefs and predictors covering demographic, education, and employment

information collected in both samples.

Digital Breadcrumbs of Migration: Data-Driven Insights into Immigrants' Online Activities

Jisu Kim, *Utrecht University*

In an era of pervasive digital connectivity, online platforms generate vast amounts of data that can provide valuable insights into the experiences and behaviors of immigrant communities. This talk presents a data-driven approach for studying immigrants and their online activities through the analysis of digital breadcrumbs derived from social media. Rather than relying on a single indicator, the approach integrates multiple signals embedded in digital traces, including linguistic patterns, cultural references, migration-related content, social network structures, geolocation information, and temporal activity patterns. By combining these diverse data dimensions, the approach enables a richer understanding of how immigrants engage, interact, and form communities in digital environments. The findings contribute to broader discussions on migration, integration, and social dynamics in the digital age, while highlighting the opportunities and challenges of using online data to study population groups.

Geocoding Uncertainty and Spatial Bias: Implications for Multi-Source Research

Manuela Schmidt, *RPTU Kaiserslautern-Landau*

Geocoding is a key step for linking survey respondents' addresses with administrative records, regional statistics, and geospatial context data. Systematic spatial errors, such as centroid placement for incomplete addresses or address sparsity in rural areas, propagate into downstream analyses that rely on geocoded locations, potentially introducing measurement error and spatial bias that vary systematically across regions.

This study presents a nationwide quantitative assessment of commonly used geocoding services in Germany, covering open-data and commercial providers. The analysis is based on two datasets: an authoritative set of hospital addresses with government-provided coordinates serving as the ground truth, and a large sample of randomly selected residential-like addresses extracted from OpenStreetMap. This design allows both accuracy assessment and relative comparison of geocoding consistency under realistic conditions, without relying on personal address data. The analysis also explores how geocoding error and provider disagreement vary with urban-rural context, population density, proximity to national borders, and address characteristics.

Building on this, the talk discusses implications for multi-source data integration: How spatial uncertainty in geocoding affects the reliability of linked contextual variables such as regional typologies, administrative unit assignments, and environmental indicators, and how it may interfere with survey-based research.



Session Info

S111

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 10:00 - 11:00 UTC

Strengthening Statistical Practice Through Collaboration: Perspectives from Women Leaders in STRATOS

USE SCHED TO ACCESS THE ZOOM ROOM

The STRATOS (STRengthening Analytical Thinking for Observational Studies) Initiative is an international collaboration of statistical and methodological experts dedicated to improving the design, conduct, analysis, and reporting of biomedical and observational research. Through accessible guidance documents, recommendations, and educational resources, STRATOS seeks to bridge methodological advances and statistical practice. Presentations illustrate diverse approaches to improving research quality, including the development of international recommendations for analyzing patient-reported outcomes in oncology trials, innovative strategies for teaching Initial Data Analysis (IDA), consensus-based development of guidance for statistical analysis plans in observational studies, and principled approaches to defining estimands and interpreting quality-of-life outcomes. These examples demonstrate how statistical thinking can improve the transparency, reproducibility, relevance, and interpretability of scientific research. Together, these talks demonstrate the impact of international collaboration, interdisciplinary research, translating complex statistical concepts into practical tools for researchers. The session also highlights the important leadership role of women statisticians in shaping methodological standards, educational initiatives, and research practices.

ORGANIZER: Marianne Huebner, Michigan State University**CHAIR:** Marianne Huebner, Michigan State University

Speaker Bios

S111



PROF. MARIANNE HUEBNER
Michigan State University

Marianne Huebner is Professor of Statistics and Director of the Center for Statistical Training and Consulting (CSTAT) at Michigan State University, USA. Her research focuses on statistical methodology for improving transparency, reproducibility, and quality in health and observational research. She has made important contributions to the field of Initial Data Analysis (IDA) and is actively involved in the development of statistical guidance and reporting recommendations through the STRATOS Initiative. In addition to methodological research, she collaborates extensively with investigators across the biomedical, public health, and sports sciences. Through her research, teaching, and mentorship, she is committed to advancing sound statistical practice and fostering the next generation of data scientists.



PROF. SASKIA LE CESSIE
Leiden University Medical Centre

Prof. Saskia le Cessie is a medical statistician working at the departments of Clinical Epidemiology and Biomedical Data Sciences of Leiden University Medical Center. Her research in medical statistics and epidemiological methods involves (1) collaboration in research projects of clinical epidemiologists and other clinical researchers in order to provide cutting edged statistical analyses in observational clinical studies and (2) performing research on statistical methodology for epidemiological studies. She is a member of the Dutch Health Council, a member of the scientific advisory board of the Dutch MS-foundation, an associate editor of Clinical Trials and a member of the Steering Group of the STRATOS initiative, which aims to provide guidance in the design and analysis of observational studies. She was work package leader of the SISAQOL-IMI work package on Single-Arm studies.



PROF. ELS GOETGHEBEUR
Ghent University Belgium

Els Goetghebeur is Professor of Statistics at Ghent University, Belgium, and a leading researcher in causal inference and biostatistical methodology. She co-chairs the Causal Inference Topic Group of the STRATOS Initiative and has contributed extensively to methodological research aimed at improving evidence generation and decision-making in health and biomedical sciences.



PROF. LARA LUSA
University of Primorska

Lara Lusa is a professor of Statistics at the Faculty of Mathematics, Natural Sciences and Information Technologies of the University of Primorska, Slovenia and Institute for Biostatistics and Medical Informatics of the University of Ljubljana, Slovenia. Her current research interests are related to the development and validation of predictive models in bio-medicine, with particular attention to applications using high-dimensional data and rare events. She is also interested in statistical programming, design of simulation studies, and teaching statistics. Beside the methodological research in statistics, she is also heavily involved as a statistical consultant to researchers in medicine.



Leading international collaborations to build consensus: Developing the SAPI Guideline within the STRATOS Initiative

Marianne Huebner, Michigan State University

The STRATOS (STRengthening Analytical Thinking for Observational Studies) Initiative is an international volunteer network of expert biostatisticians dedicated to improving the design, analysis, and reporting of observational research through practical statistical guidance and education. Its central mission is to bridge the gap between methodological developments and routine research practice by translating complex statistical concepts into accessible recommendations for analysts with diverse backgrounds.

This presentation uses the development of the Statistical Analysis Plan for Observational Studies with Initial Data Analysis (SAPI) guideline as a case study of an international Delphi consensus process. The project brought together statisticians and methodological experts from multiple countries to address the lack of guidance for statistical analysis plans in observational research and to promote transparent documentation and pre-planning of initial data analysis, data quality assessment, and analytical decision-making. Topics include defining project scope, selecting expert participants, conducting Delphi surveys, balancing methodological rigor with practical applicability, managing disagreements, and sustaining engagement in a volunteer-led international collaboration. The SAPI experience illustrates the value of international scientific networks such as STRATOS in translating methodological innovation into better research.

Developing recommendations to handle patient reported outcome data in oncology cancer trials

Saskia le Cessie, Leiden University Medical Center

In this talk I will discuss the involvement of STRATOS in the SISAQOL-IMI project. The SISAQOL-IMI Consortium aims to develop standards for designing, analysing, presenting and interpreting health-related quality of life (HRQOL) and patient-reported outcome (PRO) data in cancer clinical trials. It is an international, multidisciplinary consortium, involving regulatory agencies, health technology assessment bodies, the pharmaceutical industry, and academic and professional societies alongside experts in statistics, PRO measurement, clinical oncology, and patient advocacy. Recently, the SISAQOL-IMI consortium delivered its final recommendations (see <https://www.sisaqol.org/>). In this talk we describe these results with an emphasis of the contribution of STRATOS members and discuss some of the challenges collaborating in such an interdisciplinary consortium.

Focusing on relevant Estimands. Quality of life among the living combined with the survival comparison: a legitimate target of evidence for treatment choice.

Els Goetghebeur, Ghent University

The field of causal inference is devoted to bringing evidence on which interventions may best improve outcomes for the treated.

In an ideal world this starts from exchangeable groups, representative of a well-defined target population, which are assigned different treatments at onset. Outcomes of interest are then assessed over regular follow-up times until the study ends or patients die. How these outcomes compare at different follow-up times sheds light on what to expect under different treatment choices in target populations.

In frail populations such as late-stage oncology or in residential care centres a substantial number of patients may die during follow-up. How to interpret quality of life (QoL) after one year in the target population is then less obvious. Surprisingly, popular methods have sought to (implicitly) impute QoL after death before comparing 'augmented' results for the 'full' population that started treatment. Alternatively, composite outcomes, such as percentage with 'deteriorated QoL or death' have been reported for treatment comparisons. While the first reports appear to 'cancel' death in an attempt to avoid 'selection bias by death', the second reports no longer treat death as quite distinct from a special QoL value in the summarized endpoint. In this talk we argue that a third approach that focuses on the real-world experienced outcome in survivors, is highly relevant. It can recognize the distinct role of death and QoL-while-alive in a 2-dimensional

Teaching Initial Data Analysis: Bringing Research Advances into the Classroom

Lara Lusa, University of Primorska

The talk presents our efforts to integrate in teaching the current developments in the research conducted by the topic group on Initial Data Analysis (IDA) within the STRATOS initiative. IDA comprises all the steps between the end of data collection and the beginning of the analysis addressing the research questions. Despite its substantial impact on transparency, reproducibility, and validity, IDA is often conducted without a formal plan or is insufficiently documented. IDA is usually not thought in undergraduate and graduate courses of statistics, it is expected that this aspect of data analysis is not sufficiently understood by applied researchers. We present a case study based on the dataset that describes the characteristics of the Pokémon species, where we developed teaching activities to introduce students to effectively using IDA.



TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 12:00 - 12:30 UTC

Predicting the Right Treatment for the Right Patient: An AI-Powered Decision Support Framework Based on Predicted Individual Treatment Effects

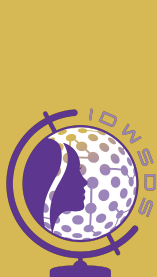
USE SCHED TO ACCESS THE ZOOM ROOM

Medical decisions—ranging from diagnosis to treatment selection—are inherently uncertain. Clinicians often rely on heuristic, experience-driven processes to integrate heterogeneous data. In this context, Predicted Individual Treatment Effects (PITE) offer a principled statistical framework to quantify how much a specific patient benefits from one treatment over another.

Advances in computational systems now enable machines to identify complex patterns within large datasets, facilitating a shift toward data-driven, individualized healthcare. This work explores using PITE to support clinical decision-making across diverse diseases and contexts. We address critical questions: which AI methods suit specific clinical datasets, how PITE should be validated, and how outcome complexity affects tool reliability.

We demonstrate PITE-based models in various disease settings, each posing unique methodological challenges. Our results show that even under real-world conditions—such as missing data—predictive models maintain interpretability and generate estimates that support clinicians. Notably, our findings highlight that internal validation is insufficient; external validation is essential for robust predictions.

Ultimately, effective PITE-based support requires more than modeling. It demands an adaptive, continuously learning system integrating data management, modeling strategies, regulatory-grade interpretability, and ongoing validation to translate evidence into precise.

ORGANIZER: Pamela Solano, Regensburg University**CHAIR:** Cherlynn Dumbura, CeSHHAR

Speaker Bios

S112



DR. PAMELA SOLANO
Regensburg University

<https://pchiroque.github.io/>

I am Pamela Solano, a statistician and researcher at the University of Regensburg, Germany. Since 2014, I have worked as a biostatistician. Following my PhD in 2018, my research focus toward statistical modeling approaches with direct societal relevance. I began working in environmental and marine biology, addressing challenges linked to environmental degradation. During this period, I rapidly established a research network and developed multidisciplinary collaborations. I was appointed as a professor in 2018 and obtained tenure in statistics in 2020. Peer-reviewed publications appear in internationally recognized journals ranked at least WoS Q2, including several Q1 outlets with impact factors above 10. My work demonstrates a commitment to society through rigorous statistical methodology, interdisciplinary reach beyond purely applied venues, and measurable societal impact in public health, climate change, environmental monitoring, and risk modeling.

Session Info

S113

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

PLENARY TALK

October 6th, 12:30 - 13:00 UTC

Navigating Data Sharing in Medical Research

USE SCHED TO ACCESS THE ZOOM ROOM

Open science is pivotal in advancing medical research by promoting accessibility and collaboration among researchers globally, and biostatisticians play a critical role in supporting these efforts. Shared datasets and software code, particularly those developed using time-intensive algorithms, are fundamental for fostering replicability and improving research efficiency for other scientists. Open access publications accompanied by publicly available data and code also help reduce disparities in knowledge access by making research materials available to researchers who might otherwise lack access. This presentation will explore data sharing principles and experiences, highlighting two recent studies: one on smoking behaviors in the United States and another on the association between caregiving and psychological well-being among Florida college students. The resulting datasets were shared under specific terms of use via Harvard Dataverse to facilitate further medical research. In addition, a subset of the tobacco use data and sample code were shared through the Resources Portal of the Teaching of Statistics in the Health Sciences Section of the American Statistical Association to support the teaching of survey methods. The presentation is intended for students, researchers, and practitioners interested in data sharing and open science.

ORGANIZER: Julia Soulakova, University of Central Florida College of Medicine

CHAIR: Cherlynn Dumbura, CeSHHAR



Speaker Bios

S113

**DR. JULIA SOULAKOVA**

University of Central Florida College of Medicine

<https://med.ucf.edu/biostatistics/>

Julia Soulakova, Ph.D., is a biostatistician and Professor of Medicine in the Department of Population Health Sciences at the University of Central Florida College of Medicine. Her research interests include statistical methodology with applications to behavioral medicine and social sciences. She is interested in examining factors associated with substance use behaviors in youth and adult populations in the United States. Her independent research has been supported by multiple external grants, including an R01 grant, funded by the NIMHD, NIH, U.S. DHHS. Her teaching portfolio includes undergraduate and graduate statistics courses. Dr. Soulakova has served as a reviewer on multiple NIH Study Sessions and was a Standing Member of the Study Session on Health Care Research Training, AHRQ, US DHHS. She is an active member of the American Statistical Association and Society for Research on Nicotine and Tobacco.



Session Info

S114

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 13:00 - 13:30 UTC

Fractional Statistical Models via Operator Theory: A Data-Driven Framework for Aviation Analytics

USE SCHED TO ACCESS THE ZOOM ROOM

Classical statistical models are built upon an assumption of short-range dependence — an assumption that fails dramatically when confronted with the complexity of real-world aviation datasets. Such datasets routinely exhibit long-range memory, non-stationarity, and heavy-tailed distributions that render conventional approaches inadequate. In this work, we propose a novel fractional statistical framework that draws on advanced operator theory to directly address these challenges, offering both rigorous theoretical guarantees and compelling empirical improvements over established baselines.

We construct a family of Toeplitz-type estimators grounded in the theory of α -fractional Bergman spaces, establish their theoretical properties, validate the framework on a large-scale aviation dataset comprising over 500,000 UAE flight records, and demonstrate prediction error reductions of 23–31% over ARIMA and 14–18% over LSTM-based approaches.

ORGANIZER: Raja'a Alnaimi, Emirates Aviation University**CHAIR:** Chiara Orsini, European Commission

Speaker Bios

S114



DR. RAJA'A ALNAIMI
Emirates Aviation University

Dr. Raja'a Al-Naimi is an Assistant Professor in the Department of Mathematics and Data Science at Emirates Aviation University (EAU), Dubai, UAE. She holds expertise in operator theory, fractional calculus, and functional analysis, with active research programs in α -fractional Bergman spaces, Toeplitz operators, and generalized numerical radius theory. Her work has been published in and accepted by leading journals including the European Journal of Pure and Applied Mathematics (EJPAM) and Taylor and Francis venues. She presented her research at the Second Sharjah International Conference on Mathematical Sciences (2026). She teaches undergraduate and graduate courses spanning multivariable calculus, analytical methods for engineers, business mathematics, statistics, and numerical methods.



Session Info

S115

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 13:30 - 14:00 UTC

Reliable Variable Selection for Biomedical Data Science: From Shrinkage Estimation to Interpretable Learning

USE SCHED TO ACCESS THE ZOOM ROOM

Modern biomedical data science increasingly relies on datasets with many predictors, limited sample sizes, and complex correlation structures. In these settings, classical regression and standard variable-selection methods may lead to unstable models, overfitting, or conclusions that are difficult to interpret. Shrinkage and penalized estimation provide a principled framework for improving reliability, but their success depends on how multicollinearity and dependence among predictors are handled. This talk discusses reliable variable selection for biomedical data science, moving from classical shrinkage ideas to modern interpretable learning. The presentation will introduce the motivation behind ridge-type methods, LASSO-based procedures, elastic net estimation, and adaptive penalization, with emphasis on approaches that use correlation information to improve model stability and interpretability. Motivated by biomedical applications such as molecular subtype identification and classification problems, the talk will show how statistically grounded regularization can support both prediction and scientific interpretation. The broader message is that reliable biomedical data science requires methods that are not only accurate, but also stable, transparent, reproducible, and interpretable for domain experts.

ORGANIZER: Mina Norouzirad, Center for Mathematics and Applications (NOVA Math)

CHAIR: Chiara Orsini, European Commission

SPONSOR: Bernoulli Society for Mathematical Statistics and Probability



Speaker Bios

S115



DR. MINA NOROUZIRAD
NOVA School of Science and Technology

<https://mnrzrad.github.io/>

Mina Norouzirad is an Assistant Researcher in Statistics at the Department of Mathematics and the Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, NOVA University Lisbon, Portugal. She is also Co-Coordinator of the Data Science Thematic Line at NOVA Math. Her research focuses on statistical learning, shrinkage and penalized estimation, high-dimensional data analysis, and biomedical applications. She is the author of Rank-Based Methods for Shrinkage and Selection with Application to Machine Learning and co-editor of Advances in Shrinkage and Penalized Estimation Strategies. She has also contributed book chapters, peer-reviewed journal articles, and R software packages, combining methodological research with reproducible computational tools. She is active in scientific service and event organization.



CAREER OR PROFESSIONAL DEVELOPMENT SESSION

PLENARY TALK

October 6th, 14:00 - 14:30 UTC

Data, Equity and Power: Institutional Frameworks for Embedding Young African Women in Decision-Making

USE SCHED TO ACCESS THE ZOOM ROOM

Despite advances in data science and statistical training across Africa, a structural disconnect persists between the development of technical talent and its integration into national and regional policy processes. It is particularly pronounced for early-career African women statisticians, who face intersecting institutional barriers. UN Women and PARIS21 shows that women occupy only 15% of chief statistician positions in Sub-Saharan Africa, reflecting patriarchal institutional cultures, limited senior-level mentorship and rigid career progression pathways.

This abstract proposes an actionable policy framework to transition young African women statisticians from technical implementers to strategic policy influencers. Using a comparative case-study methodology through the Young African Statisticians Association network, we examine institutional barriers and entry pathways across selected National Statistical Offices in East and West Africa.

The framework advances three strategic pillars: institutional quotas and fast-track leadership pathways for young women in state-led data initiatives; coordinated mentorship systems linking global statisticians with local networks to institutionalize peer and senior sponsorship and gender-responsive data governance through dedicated advisory positions for young women statisticians in ministerial policy processes.

Embedding young women statisticians in Africa's governance is vital for equitable, evidence-based development.

ORGANIZER: Sarah Nzioka, RefugePoint, MEL Manager**CHAIR:** Qi Deng, York University

Speaker Bios

S116



MS. SARAH NZIOKA

Young Africa Statisticians Association

<https://youngstatisticians.org/yasa-kenya/>

A results-oriented MEL Management professional with 8+ years of progressive multi-sector experience in monitoring, evaluation, and managing complex project lifecycles from inception to completion. Possessing proven expertise in designing and implementing tailored M&E & research frameworks, tools and methodologies, currently serve as a lead for strategic research and development programs at YASA.

I leverage my technical background to drive institutional equity and structural policy transformation across Africa. I specialize in utilizing evaluative frameworks like Outcome Harvesting to track behavioral and institutional shifts within national data ecosystems. Actively spearhead YASA's regional advocacy initiatives to dismantle systemic barriers and establish gender-transformative leadership pipelines that embed early-career women statisticians directly into National Statistical Offices (NSOs)



Session Info

S117

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 14:30 - 15:00 UTC

Privacy Doesn't End at the Match: Querying PPRL Data in Practice

USE SCHED TO ACCESS THE ZOOM ROOM

Privacy-preserving record linkage (PPRL) has a rich methodological literature on encoding, matching, and cryptographic guarantees — but comparatively little guidance exists on what happens after the match: how do researchers actually access, query, and analyze the linked data that results? This presentation addresses that gap directly, walking through real-world architectures used by statistical agencies, such as the U.S. Census Bureau's linkage key infrastructure and UK Trusted Research Environments, to control researcher access to linked data. We'll cover how to handle potential uncertainty when analyzing linked data, such as producing confidence tiers, adjusting thresholds, and a novel approach to clerical review within the PPRL system. We'll also walk through concrete querying structures such as linkage maps and secure enclaves and will explore examples of queries and their outputs. Attendees will leave with a clearer picture of the access-and-analysis landscape for linked data, practical considerations for designing their own linkage projects, and a better sense of how privacy shapes analysis.

ORGANIZER: Emily Gentles, RTI International**CHAIR:** Mina Norouzirad, Center for Mathematics and Applications (NOVA Math)

Speaker Bios

S117



EMILY GENTLES
RTI International

Emily Gentles is an expert in data linkage, including entity resolution and privacy-preserving record linkage (PPRL). Ms. Gentles has researched efficient PPRL methods; worked to develop secure linkage systems; and designed innovative record linkage procedures, including manual review protocols. She graduated from Duke University with a Masters in Statistical Science in 2022 and now works at Research Triangle Institute as a Research Statistician.



Session Info

S118

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 15:00 - 16:00 UTC

Data, Diversity, and Dialogue: Women Shaping the Future of Statistics and Data Science

USE SCHED TO ACCESS THE ZOOM ROOM

At a time when data drives decisions across science, industry, and society, the voices shaping this landscape matter more than ever. This invited session brings together distinguished women researchers whose work reflects the power of diversity, collaboration, and intellectual curiosity in statistics and data science.

Through their journeys, the speakers will illustrate how meaningful connections—across disciplines, institutions, and cultures—emerge not only from shared research goals but also from openness, mentorship, and resilience. Their experiences reveal how collaborative environments can spark innovation, amplify impact, and create pathways for future generations of women in the field.

More than a showcase of research, this session is an invitation to reflect on how we build scientific communities that are inclusive, dynamic, and forward-looking. By fostering dialogue and embracing diverse perspectives, we strengthen our collective ability to tackle complex challenges and shape a more connected and equitable future for data science.

This session celebrates the role of women as catalysts for change, bridging ideas and people to advance knowledge and inspire the next generation.

This session is sponsored by the Portuguese Statistical Society and by the Portuguese Country Representative of CWS.

ORGANIZER: Eunice Carrasquinha, Faculdade de Ciências da Universidade de Lisboa and CEAUL

CHAIR: Eunice Carrasquinha, Faculdade de Ciências da Universidade de Lisboa and CEAUL



Speaker Bios

S118



DR. CONCEIÇÃO AMADO
Instituto Superior Técnico and CEMAT

Conceição Amado is an Associate Professor in the Department of Mathematics at Instituto Superior Técnico, University of Lisbon. She is a researcher at the Center for Computational and Stochastic Mathematics (CEMAT) and serves as co-coordinator of the M.Sc. in Data Science and Engineering and the Ph.D. in Statistics and Data Mathematics. Her research interests include robust multivariate analysis and statistical methods for data mining, encompassing both methodological developments and practical applications, particularly in complex and high-dimensional data settings. She has been involved in several national and international research projects.



DR. MARÍLIA ANTUNES
Universidade de Lisboa and CEAUL

Marília Antunes is an Associate Professor at the Department of Mathematical Sciences at the Faculty of Sciences, University of Lisbon. Since 2023, she is also the Scientific Coordinator of the Centre of Statistics and its Applications (CEAUL). She earned her PhD in Statistics and Operations Research in 2003 and since then has been mostly dedicated to Biostatistics, particularly its applications to Biology, Medicine, and Neuroscience, leading to active collaborations with researchers in these fields. She is actively involved in national and international research projects and currently serves as the local PI of projects such as CINDERELLA and ResPECT, where AI-driven methodologies are developed to support more personalized, efficient, and equitable healthcare. These projects reflect her enthusiasm and commitment to creating global scientific connections and fostering multidisciplinary collaboration across institutions and countries.



DR. LISETE SOUSA
Universidade de Lisboa and CEAUL

Lisete Sousa is an Associate Professor in the Department of Mathematical Sciences at the Faculty of Sciences of the University of Lisbon and a researcher at the Centre of Statistics and its Applications (CEAUL). She holds a degree in Statistics and Operations Research, a master's degree in Probability and Statistics, and a PhD in Probability and Statistics, all from the Faculty of Sciences of the University of Lisbon (FCUL). Her research focuses on the statistical analysis of high-dimensional omics data, addressing challenges such as normalization, feature selection, and classification, with applications in biology, biomedicine, and public health. Lisete Sousa served as Director of the Master's in Biostatistics and Director of CEAUL from 2017 to 2022. Currently, she is Vice-Director of CEAUL, Vice-President of the Portuguese Statistical Society (SPE), and Associate Editor of REVSTAT - Statistical Journal.

<https://web.ist.utl.pt/susanavinga/>



DR. SUSANA VINGA
INESC-ID, Instituto Superior Técnico and IDMEC

Susana Vinga is Associate Professor at the Departments of Computer Science and Engineering and Bioengineering at Instituto Superior Técnico, Universidade de Lisboa, and researcher at INESC-ID (Life and Health Technologies). She holds a degree in Mechanical Engineering (1999), a post-graduate degree in Probability and Statistics (2002), a PhD in Biology/Bioinformatics (2005), and a Habilitation in Computer Science and Engineering (2025). Her research bridges computer science, engineering and life sciences, focusing on systems and computational biology, dynamic modeling, control and optimization of biological networks, and machine learning/biostatistics for personalized medicine. She has led major funded projects, including the H2020 Twinning OLISSIPO and several national projects. She has over 100 ISI-indexed publications (h-index=26) and multiple distinctions, including the ULisboa/CGD Scientific Prize (2017, 2022) and UTL Young Researcher Award (2010).



Breaking the Zero Wall: Robust Poisson Hurdle Models*Conceição Amado, Instituto Superior Técnico and CEMAT*

Many real-world problems involve count data with excess zeros, overdispersion and outliers, where standard Poisson or negative binomial models perform poorly. Typical examples arise in health-related outcomes (such as the number of adverse events per patient or per facility), in industrial quality monitoring (for instance, defect counts in steel manufacturing processes), and in the analysis of rare events per unit of time. Hurdle models address these challenges by separating the binary mechanism that governs whether any event occurs from the process that generates strictly positive counts. In this talk, I present a robust estimation approach for the Poisson hurdle model based on minimum distance methods. The procedure, built on distances such as the Hellinger and total variation distances, is applied separately to the binary and truncated Poisson components, preserving the interpretability of the two-part structure. A simulation study inspired by applications with many zeros and atypical observations shows that the minimum distance estimators maintain good efficiency in clean samples while substantially improving robustness relative to maximum likelihood and other existing estimators. The methodology will also be illustrated with practical examples, including industrial defect data, highlighting how robust hurdle modelling can support real-world decision-making in diverse domains.

Uncovering True Engagement: A Mixture Model for Active vs. Unattended Sessions*Marília Antunes, Faculdade de Ciências da Universidade de Lisboa and CEAUL*

The CINDERELLA project aims to improve aesthetic outcomes and patient satisfaction following breast cancer surgery through an AI-based application that predicts post-surgical aesthetic results from patient data and medical images. App usage data provides an opportunity to study user behaviour; however, a software bug caused incorrect recording of session times. Recovering durations from login data may generate inflated values between sessions, while additional inflation may occur when users leave the app open without active interaction. Thus, observed usage times may arise from two latent processes: active and inactive periods.

A two-component finite mixture model is used to classify session times into exact or inflated categories. Log-normal and Gamma specifications were explored, incorporating socio-demographic, clinical, and user-application interaction covariates through mixtures of generalized linear models. Model estimation is performed using EM algorithms in a maximum likelihood framework and NUTS algorithms in a Bayesian framework.

From Data to Dialogue: Building Connections in Women-Centred Statistics and Data Science*Lisete Sousa, Faculdade de Ciências da Universidade de Lisboa and CEAUL*

Statistics and data science are often presented as technical

disciplines driven by methods, algorithms, and computational power. Yet, behind every dataset, there are people, stories, and communities that shape how knowledge is created and shared. In this talk, I reflect on my journey as a statistician and researcher working across interdisciplinary collaborations in health, life sciences, and bioinformatics, highlighting the role that dialogue, mentorship, and scientific generosity have played throughout that path.

Rather than focusing solely on methodological advances, this presentation explores how meaningful scientific work is frequently built through collaborative environments where diverse perspectives are valued and encouraged. I will discuss how women researchers and mentors have influenced my own scientific identity and how these experiences continue to shape my approach to teaching, supervision, and research collaboration.

The talk will also briefly present current work related to microbiome research and menopause-associated health risks, a project combining statistical learning, bioinformatics, and women's health data to better understand complex biological and social dimensions of ageing.

By sharing these experiences, I hope to contribute to a broader conversation about visibility, inclusion, and the importance of supportive scientific communities that empower future generations of women in statistics and data science.

Interpretable Machine Learning for Prognosis: When Sparsity Meets Biological Networks*Susana Vinga, INESC-ID, Instituto Superior Técnico and IDMEC*

Many biomedical prediction tasks, such as tumor classification, relapse risk, and other time-to-event outcomes, are dominated by the “small n , large p ” regime, where multiple models can fit the data equally well, and naïve feature selection becomes unstable. In this talk, I discuss how sparse learning and structured regularization help turn ill-posed problems into actionable and reproducible knowledge. I will introduce regularized generalized linear models and Cox regression enriched with network information, such as gene centrality, correlation structure and smoothness over biological graphs, to obtain compact and robust biomarker panels with clearer mechanistic interpretation. Drawing on cancer case studies, I will illustrate how network-aware penalties can improve generalization, reduce signature variability, and support clinically meaningful risk stratification. I will also emphasize practical validation lessons for high-dimensional biomedical data, including repeated resampling, stability analysis and transparent reporting of uncertainty. The goal is to bridge statistics, optimization and life sciences to build models that remain interpretable enough to support scientific insight and, ultimately, clinical decision support in precision medicine.



Session Info

S119

TECHNICAL STATISTICAL SESSION

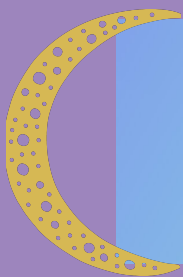
CONTRIBUTED SESSION

October 6th, 17:00 - 18:00 UTC

Statistical Methods and Applications 3

USE SCHED TO ACCESS THE ZOOM ROOM

Adaptive Functioning in Angelman Syndrome: Insights from Statistical Modeling | Novel Influence Diagnostics in Multistate Models for Breast Cancer | Goodness-of-fit test for the Dirichlet distribution | A Bayesian Exploration of Bimodal Fertility Patterns Using Lifetime Densities.

ORGANIZER: IDWSDS Program Committee**CHAIR:** Sima Sharghi, Akron Children's Hospital

Speaker Bios

S119



MS. JULIA ZHANG
RTI International

Julia Ann Zhang graduated from North Carolina State University with a Masters Degree in Statistics and a Concentration in Biostatistics. Currently, she is a statistician at Research Triangle Institute (RTI), where she provides data processing, statistical programming, and data analysis in both R and SAS to support a variety of projects across domains, such as health, energy/environment, and education/human development. She has also worked as an Advanced Biostatistics Intern at United Therapeutics, analyzing clinical trial data, as well as a graduate research assistant at North Carolina State University, where she helped develop a dynamic map and dashboard in R Shiny to present descriptive statistics of PFAS distribution across geographic regions in North Carolina. In her free time, Julia enjoys running, yoga, and cooking.

<https://github.com/Valeria-Leiva>



DR. VALERIA LEIVA
Universidad de los Andes, Chile

I am a Statistician with B.Sc., M.Sc., and Ph.D. training in Statistics from Pontificia Universidad Católica de Chile. I am currently affiliated with the Department of Public Health at Universidad de los Andes, Chile. My research focuses on joint modeling of longitudinal and survival data, with particular emphasis on multistate models and their application to breast cancer research. I completed a postdoctoral fellowship at the MRC Biostatistics Unit, University of Cambridge, where I specialized in the development and application of multistate models for breast cancer studies. My broader research interests include statistical methodology for public health, survival analysis, genetic statistics, and computational methods for health research.

<https://www.linkedin.com/in/sucharitha-dodamgodage/>



MS. SUCHARITHA DODAMGODAGE
Clarkson University

Sucharitha (Suchi) Dodamgodage is a Biostatistician and Statistics PhD Candidate at Clarkson University who loves turning complex mathematical theory into clean, reliable data pipelines for the life sciences. She solved a major statistical conjecture left open since 1989 regarding Dirichlet distribution estimators, transforming that breakthrough into an end-to-end pipeline featuring parameter estimation, differential abundance analysis, and goodness-of-fit frameworks for noisy, high-dimensional microbiome data. Beyond her dissertation, her work spans epidemiology and pediatric health. Suchi pairs this deep theoretical background with hands-on corporate experience as a biostatistician intern at Inari Medical in California and Vironix Health in New York. A natural leader who served as Chapter President and a mentor for the Association for Women in Mathematics (AWM). Set to graduate in Fall 2026, Suchi is currently open to full-time industry roles where she can make an immediate impact.

<https://shambhavisingh010.github.io/>



DR. SHAMBHAVI SINGH
University of Petroleum and Energy Studies, Dehradun

I am working as an Assistant Professor (Statistics), Applied Sciences Cluster at the School of Advanced Engineering, University of Petroleum and Energy Studies, Uttarakhand, India. I have obtained my Ph.D. in Statistics (2025) from Banaras Hindu University under the supervision of Prof. Satyanshu Kumar Upadhyay. During my PhD, I received the DST-INSPIRE fellowship. My research interests include Bayesian statistics and computation, demographic modeling, and reliability analysis.



Adaptive Functioning in Angelman Syndrome: Insights from Statistical Modeling*Julia Zhang, RTI International*

Modeling person ability scores in rare disease natural history studies presents several methodological challenges, including nonlinear developmental trajectories, heterogeneous variance, and floor and ceiling effects. This presentation describes a modeling framework developed to address these challenges using data from a large, multi-site longitudinal study of individuals with Angelman syndrome (AS).

AS is a rare neurogenetic condition characterized by severe to profound intellectual disability, minimal speech, and motor impairments. We focus on modeling adaptive functioning over age, using Vineland-3 Growth Scale Value scores across multiple subdomains. To capture the complex structure of these data, we evaluated and compared a series of candidate models and selected an approach that balances interpretability with statistical rigor while accommodating key features of the data, including nonlinearity and heteroscedasticity.

We present the model development process, including comparison of alternative specifications, and highlight how the selected framework captures distinct developmental trajectories across molecular subtypes of AS. Results demonstrate meaningful differences in patterns of adaptive functioning by subtype and illustrate how model choice influences substantive conclusions. We conclude with practical recommendations for modeling longitudinal person ability scores in rare disease contexts, with implications for study design, analysis, and interpretation.

Novel Influence Diagnostics in Multistate Models for Breast Cancer*Valeria Leiva, Department of Public Health, Universidad de los Andes, Chile*

Multistate models were developed to model survival data where several midpoints and endpoints are of interest; and they have been particularly successful in modeling dynamics of cancer. As in any statistical model, identification of influential observations is an essential task, as they can significantly affect the validity of inferred parameters and conclusions drawn from the data. The local influence approach is a set of methods designed to detect the effect of small perturbations of the model or data on the inference, allowing for deeper data analysis. In this paper, we derive local influence methods for multistate models and illustrate their use with a breast cancer dataset. In particular, we develop and implement local influence diagnostic techniques based on a suitable estimation equation. For simplicity, we restrict our consideration to the Multistate Proportional Hazards model, using different case-weight perturbation strategies.

lacks a rigorous goodness-of-fit (GoF) framework to validate its assumptions. Existing methodologies face computational barriers and lack documented performance for dimensions $d > 2$. This study evaluates the implementation, test size, and Power of standard Empirical Distribution Function (EDF) tests, specifically Kolmogorov-Smirnov (KS), Anderson-Darling (AD), and Cramér-von Mises (CvM), alongside a novel Parametric Bootstrap Chi-square (PB-CS) test. Implementing EDF tests for the Dirichlet distribution introduces two major hurdles: the absence of a closed-form distribution function for $d > 2$ and the mathematical ambiguity of ranking multivariate observations. To bypass the closed-form limitation, we propose a two-step transformation using Beta decomposition and the Rosenblatt transformation to map $(d-1)$ -dimensional unit simplex data into a $(d-1)$ -dimensional unit hypercube. Furthermore, we demonstrate that the frequency-based PB-CS test completely bypasses the multivariate ranking problem. To adapt it for continuous data, we introduce an extended Doane's rule for bin selection paired with a quasirandom binning strategy using Sobol sequences and Voronoi partitioning. Extensive simulations show that while EDF methods thrive in lower dimensions, the PB-CS framework delivers superior or comparable performance as dimensionality increases.

A Bayesian Exploration of Bimodal Fertility Patterns Using Lifetime Densities.*Shambhavi Singh, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India.*

The work explores mixtures of three lifetime densities, viz., normal, Weibull, and generalized gamma densities, to capture fertility patterns characterized by hump- or bimodal-shaped behavior, an intriguing trend that researchers have observed over the last two to three decades. It harnesses the Bayesian framework to achieve a comprehensive posterior analysis. The Metropolis algorithm was used with a carefully selected candidate-generating density to produce posterior samples for each case. Numerical illustrations have been conducted using unabridged real-life data. Ultimately, the proposed models were compared to identify the most suitable model, ensuring compatibility with the available data. This work is inspired by previous studies that have largely overlooked the use of more flexible lifetime models in analyzing fertility patterns of this nature, prompting us to incorporate them into our analysis.

Goodness-of-fit test for the Dirichlet distribution*Sucharitha Dodamgodage, Clarkson University*

The Dirichlet distribution is the first choice for modeling compositional data like microbiome profiles, yet the literature



Session Info

S120

TECHNICAL STATISTICAL SESSION

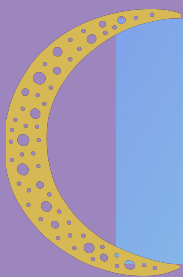
INVITED SESSION

October 6th, 18:00 - 19:00 UTC

Statistical Modeling for Applied Research in Costa Rica

USE SCHED TO ACCESS THE ZOOM ROOM

This session brings together four applied studies demonstrating how modern statistical methods can address diverse scientific questions using complex data from Costa Rica. The presentations cover Bayesian hierarchical modeling, penalized regression, machine learning, and generalized additive models, with applications to infectious disease surveillance, biotechnology awareness, and animal production. Despite spanning different disciplines, the talks share a common emphasis on developing interpretable and robust statistical approaches to better understand complex systems through real-world applications in the Costa Rican context.

ORGANIZER: Shirley Rojas, University of Costa Rica**CHAIR:** Shirley Rojas, University of Costa Rica

Speaker Bios

S120



MISS JIMENA MURILLO-MONTERO
Boston Scientific

I grew up in Costa Rica, in a family with strong academic roots. From an early age, I was drawn specially to mathematics and biology. In 2017, I began studying Microbiology and Clinical Chemistry at the University of Costa Rica. In 2018 I added Statistics as a second major, but during the pandemic, I decided to focus on Statistics and graduated in 2023 with a final project on dengue fever prediction using neural networks. I later worked as an instructor at the School of Statistics at UCR, teaching service courses during 2023 and 2024, mostly for health sciences students.

At the end of 2024, I joined a medical device manufacturing company, as a Data Scientist. My work focuses on material traceability using graph models, regression, machine learning and other tools to support manufacturing problem-solving.

I am currently pursuing a master's degree in Statistics, developing a thesis on dengue fever underreporting and statistical methods to better understand disease surveillance data.



MISS LAURA OBANDO ESQUIVEL
Universidad de Costa Rica

Laura Obando Esquivel is a student in the Bachelor's Degree in Statistics at the University of Costa Rica and a researcher at the School of Biology, working with the Plant Biotechnology Laboratory and the Center for Research in Pure and Applied Mathematics at the same institution. In her free time, she enjoys reading, poetry, and cinema.



MISS MARIANNE PEÑA WÜST
University of Costa Rica

Marianne is a senior Statistics student from the University of Costa Rica with a background in competitive mathematics. Currently, she has been a research assistant at the Centro de Investigación en Matemática Pura y Aplicada (CIMPA), where she collaborates with projects related to environmental data science, particularly agriculture and hydrology. She is also a climate activist and has been a part of Fridays for Future Costa Rica and the Frente Ecologista Universitario (FECOU).



DR. SHIRLEY ROJAS-SALAZAR
University of Costa Rica

Shirley Rojas is a professor at the School of Statistics, University of Costa Rica. She completed her undergraduate and graduate studies in Statistics. She enjoys applying statistical methods to better understand complex biological processes and address questions in agricultural and animal production systems.



Bayesian Spatio-Temporal Count Models with Underreporting: A Practical Identifiability Approach for Dengue Surveillance in Costa Rica

Jimena Murillo-Montero, Boston Scientific

Underreporting is a major challenge in epidemiological surveillance because reported cases do not necessarily reflect the true incidence of disease. This issue is particularly relevant for dengue in Costa Rica, where surveillance relies on passive reporting and laboratory confirmation, potentially biasing spatio-temporal analyses and public health decision-making. This study proposes Bayesian spatio-temporal count models with explicit underreporting to jointly estimate latent disease incidence and reporting probability. The primary objective is to evaluate the practical identifiability of these models, that is, their ability to reliably distinguish the disease process from the reporting process using finite observational data. Poisson-logit and Negative Binomial-logit specifications incorporating spatial and temporal random effects will be compared through simulation studies designed to assess parameter recovery, estimation stability, and the contribution of model structure to practical identifiability. The selected model will subsequently be applied to weekly dengue surveillance data from Costa Rica (2017–2025). Simulation and application results are expected to illustrate the performance of the proposed framework and its potential to improve estimation of dengue burden under underreporting, providing more reliable evidence for epidemiological surveillance and public health decision-making.

The gender gap in biotechnology awareness: a ridge regression analysis from Costa Rica

Laura Obando Esquivel, Student

Biotechnology awareness varies considerably across populations, yet the factors driving this variation remain poorly characterized in Costa Rica. This study examines sociodemographic and attitudinal predictors of self-reported biotechnology awareness among Costa Rican adults using data from the 2022 Encuesta de Actualidades (Current Affairs Survey, University of Costa Rica, $n = 1,744$). Ridge regression (penalized logistic regression, $\alpha = 0.0014$) was applied to address multicollinearity among predictors while retaining all variables in the model. Willingness to consume genetically modified foods (OR = 1.78), educational attainment (OR = 1.64), and reading capacity (OR = 1.44) showed the strongest positive associations with awareness. Being a woman was associated with lower odds of biotechnology awareness (OR = 0.68). Perceived biotechnology knowledge (OR = 1.35) and household income (OR = 1.08) showed smaller positive associations with awareness. These findings suggest that science communication strategies targeting women and populations with lower educational attainment may be central to reducing biotechnology awareness gaps in Costa Rica.

Dengue fever is a disease that poses a significant health risk in tropical and subtropical regions. Therefore, this study aims to predict weekly dengue relative risk for 82 municipalities in Costa Rica from 2020 to 2025, by using meteorological and social variables and Google Trends data. The modeling approach included machine learning techniques such as Random Forest, XGBoost, and their hurdle variants to account for excess zeros. Each of the four models was fitted independently for each municipality, in addition to a global model fitted using data from all municipalities. Performance was evaluated using Normalized Root Mean Squared Error, and variable importance and selection were computed. Random Forest models exhibited the best performance. The most important variables included relative risk lags, meteorological variables, and Google Trends topics relating to employment, mobility and finance, and dengue spread.

Modeling Milk Yield in a Tropical Dairy Herd with Generalized Additive Models

Shirley Rojas-Salazar, University of Costa Rica

Milk yield is influenced by biological factors that often exhibit nonlinear relationships, making conventional linear models inadequate to describe lactation dynamics. This study applied Generalized Additive Models (GAMs) to evaluate the effects of days in milk, parity, and breed composition on milk yield in a tropical dairy herd. Data were collected at the Diamante animal production farm in Guápiles, Costa Rica. Smooth functions were used to model nonlinear effects, and cow was included as a random effect to account for repeated measurements. The analysis identified nonlinear effects of the explanatory variables and enabled the comparison of alternative model specifications. The results demonstrate the usefulness of GAMs for modeling complex biological relationships and providing an interpretable framework for analyzing longitudinal milk production data.

Forecasting Weekly Dengue Relative Risk Across Costa Rican Municipalities Using Machine Learning and Digital Surveillance Data

Marianne Peña Wüst, Student



Session Info

S121

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 20:00 - 21:00 UTC

Advances in Statistical Methodology for Complex and High-Dimensional Data

USE SCHED TO ACCESS THE ZOOM ROOM

This invited session showcases recent advances in statistical methodology for analyzing increasingly complex, high-dimensional, and structured data arising in biomedical research, imaging, surveys, and longitudinal studies. The presentations introduce innovative approaches to sufficient dimension reduction, scalable spatial topic modeling for multiplexed imaging, Bayesian small area estimation under informative sampling, and Bayesian semiparametric quantile mixed-effects models for longitudinal data with missing responses. Together, these talks highlight modern statistical techniques that combine methodological rigor with computational efficiency to address emerging challenges in data science, while demonstrating broad applicability across diverse scientific disciplines.

ORGANIZER: Yet Nguyen, Old Dominion University**CHAIR:** Yet Nguyen, Old Dominion University

Speaker Bios

S121



DR. MAI DAO
Wichita State University

https://www.wichita.edu/profiles/academics/fairmount_college_of_liberal_arts_and_sciences/Math/Faculty/Dao-Mai.php

Dr. Mai Dao is an Assistant Professor of Statistics in the Department of Mathematics, Statistics, and Physics at Wichita State University. She obtained her Ph.D. from Texas Tech University under the supervision of Professors Min Wang (The University of Texas at San Antonio) and Souparno Ghosh (University of Nebraska-Lincoln). Her primary research interests focus on Bayesian statistics, high-dimensional inference, and statistical machine learning.



DR. CHENLU KE
Virginia Commonwealth University

<https://ssor.vcu.edu/directory/ke.html>

Dr. Chenlu Ke is an Associate Professor in the Department of Statistical Sciences and Operations Research at Virginia Commonwealth University. Her research focuses on methodological and computational developments for high- and ultrahigh-dimensional complex data, with applications in biomedical sciences. In particular, she has been working on sufficient variable selection, sufficient dimension reduction, and related explainable statistical learning methods for extracting low-dimensional structure from complex data and outcomes.



DR. YANGHYEON CHO
University of Idaho

<https://www.uidaho.edu/people/ycho>

Yanghyeon Cho is an Assistant Professor in the Department of Mathematics and Statistical Science at the University of Idaho. Her research interests include survey sampling and statistical genetics, with a focus on small area estimation, Mendelian randomization, and multi-omics data integration, with applications to agricultural surveys and Alzheimer's disease.



DR. XIYU PENG
Texas A&M University

<https://sites.google.com/view/xiyupeng/home>

Dr. Xiyu Peng is an assistant professor in the Department of Statistics at Texas A&M University. Her research focuses on developing statistical and computational methods for resolving temporal and spatial dynamics in single cell omics data with applications in cancer. She got her Ph.D. from Iowa State University and did postdoc training in Memorial Sloan Kettering Cancer center.



Bayesian Quantile Semiparametric Mixed-Effects Double Regression Models for Analyzing Longitudinal Data with Non-Ignorable Missing Responses

Mai Dao, *Wichita State University*

Semiparametric mixed-effects double regression models have demonstrated satisfactory efficacy across diverse applications of longitudinal studies. However, the estimation of these models often relies on the assumptions of normally distributed errors and complete data with no missing values. Therefore, such restrictions may limit the practical usage of these models when analyzing longitudinal data that exhibit heavy-tailed behaviors and/or contain missing values. In this talk, we introduce a Bayesian quantile regression-based semiparametric mixed-effects double regression model for examining longitudinal data with non-ignorable missing responses. Here, the quantile regression is used to address non-normality issues and the missing mechanism is defined through a logistic regression model. Our proposed algorithm can concurrently model both the mean and variance of the mixed effects as functions of predictors while investigating the covariate effects at different quantiles of interest. Additionally, we utilize the Bayesian adaptive LASSO hierarchical model to devise an effective Metropolis-Hastings-within-Gibbs computation algorithm for both parameter estimation and variable selection purposes. Finally, we conduct different simulation studies and a real-data example to demonstrate the successful implementation of our proposed Bayesian methodology.

Explained Kernel Embedding Variation for Sufficient Dimension Reduction

Chenlu Ke, *Virginia Commonwealth University*

We propose a new framework for sufficient dimension reduction based on explained kernel embedding variation. The main idea is to view dimension reduction through a model-free ANOVA decomposition in reproducing kernel Hilbert spaces and to estimate low-dimensional linear combinations of predictors that maximize regression information. This formulation unifies several SDR targets: with characteristic kernels it targets the central subspace, while with linear kernels it reduces to central mean subspace estimation. On the computational side, we combine a sequential procedure that builds an ordered set of directions with a joint refinement step over the Grassmann manifold. We also develop a cross-validation-based dimension selection procedure using held-out kernel embedding prediction loss. The proposed method is flexible, model-free, and naturally accommodates different response kernels depending on the inferential target. Numerical studies illustrate its ability to recover both mean and distributional sufficient directions, and comparisons with existing sufficient dimension reduction methods demonstrate its competitive performance.

Comparison of Small Area Procedures Based on Gamma Distributions with Extension to Informative Sampling

Yanghyeon Cho, *University of Idaho*

The gamma distribution is a useful model for small area

prediction of a skewed response variable. We study the use of the gamma distribution for small area prediction. We emphasize a model, called the gamma-gamma model, in which the area random effects have gamma distributions. We compare this model to a generalized linear mixed model. Each of these two models has been proposed independently in the literature, but the two models have not yet been formally compared. We evaluate the properties of two mean square error estimators for the gamma-gamma model, both of which incorporate corrections for the bias of the estimator of the leading term. Finally, we extend the gamma-gamma model to informative sampling. We conduct thorough simulation studies to assess the properties of the alternative predictors. We apply the proposed methods to data from an agricultural survey.

A Scalable Bayesian Spatial Topic Model for Large-Scale Multiplexed Imaging Analysis

Xiyu Peng, *Texas A&M University*

Recent progress in multiplexed tissue imaging is deepening our understanding of tumor microenvironments related to treatment response and disease progression. However, analyzing whole-slide images with millions of cells remains computationally challenging, and few methods provide a principled approach for integrative analysis across images. Here, we introduce SpatialTopic, a spatial topic model designed to decode high-level spatial tissue architecture from multiplexed images. This algorithm integrates both cell type and spatial information within a topic modelling framework, originally developed for natural language processing and adapted for computer vision. The model employs an efficient collapsed Gibbs sampling algorithm for inference. We benchmarked SpatialTopic across diverse single-cell spatial transcriptomic and proteomic imaging platforms spanning multiple tissue types. We show that SpatialTopic is highly scalable to large-scale images, along with high precision and interpretability. It consistently identifies biologically and clinically significant spatial topics, such as tertiary lymphoid structures, and tracks spatial changes over disease progression. Its computational efficiency and broad applicability will enhance the analysis of large-scale imaging datasets.



Session Info

S122

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 21:00 - 22:00 UTC

Advances from Junior Bayesian Statisticians: Bayesian Models for Complex Networks and Relational Data

USE SCHED TO ACCESS THE ZOOM ROOM

The junior section of the International Society of Bayesian Statistics (j-ISBA) is excited to present a session showcasing the work of outstanding early-career women researchers from the United States, Italy, and Germany. This session highlights recent advances in Bayesian methodology for analyzing complex relational and structured data, with a focus on flexible probabilistic modeling, scalable computation, and impactful real-world applications.

The four talks highlight modern Bayesian approaches for the modeling of networks and high-dimensional relational data. They span generative models for graph-structured data grounded in classical Bayesian principles of exchangeability, Bayesian nonparametric models for interaction networks with evolving community structure, latent variable models that account for uncertainty and reporting bias in social network data, and hierarchical tensor factorization methods for large-scale sparse relational arrays. Across these contributions, speakers develop computationally efficient inferential strategies using tools including variational inference, hybrid Monte Carlo algorithms, and scalable latent variable methods. These approaches enable Bayesian analysis in settings involving millions of observations while maintaining interpretable probabilistic models.

Together, these talks demonstrate the innovative methodological, theoretical, and applied contributions of early-career women Bayesian statisticians.

ORGANIZER: Emma Landry, Johns Hopkins University

CHAIR: Emma Landry, Johns Hopkins University



Speaker Bios

S122



DR. FEDERICA ZOE RICCI
Swarthmore College

<https://federiczoe.github.io/>

Federica is an Assistant Professor of Statistics (tenure-track) at Swarthmore College, a liberal-arts college located near Philadelphia. She received her PhD in Statistics from UC Irvine. She is currently interested in the development of modeling approaches and scalable computational methods for network analysis, combining Bayesian statistics and machine learning. Federica is equally passionate about teaching, and she works on the development of game-based activities for teaching statistics, on understanding the emergent trends in data science education and on an R package for grading called gradetools.



DR. LOUISE ALAMICHEL
Bocconi University

<https://alamichl.github.io/>

Louise is a post-doctoral researcher at Bocconi University in Milan, working with Daniele Durante. Her research is on Bayesian non-parametric mixture models. On one side, she focuses on the inference of network data using these models, and on the other side, on their asymptotic properties for the clustering of data.



DR. MARTINA CONTISCIANI
Goethe University Frankfurt

<https://mcontisc.github.io/>

Martina Contisciani is a postdoctoral researcher at the Center for Critical Computational Studies (C3S) at Goethe University Frankfurt. Her research focuses on the development of statistical models and algorithms for the study of complex systems, with a particular interest in inferential methods. She is also passionate in applying these methodologies to real-world problems and she is currently involved in the MOMA project, which aims to model the multiscale network structure of migration flows in Austria. Martina received her Ph.D. in Computer Science in 2024 from the Max Planck Institute for Intelligent Systems and the University of Tübingen. She also holds a master's degree in Data Science and a bachelor's degree in Statistics for Business and Economics, both from the University of Padua.



DR. JIE JIAN
The University of Chicago

<https://sites.google.com/view/jiejian/home>

Jie is a postdoctoral scholar at the Data Science Institute at the University of Chicago. Her research develops interpretable probabilistic and machine-learning methods for network and tensor data. These methods are motivated by applications in brain imaging, climate science, international trade and relations, and linguistics. Jie completed her PhD in Statistics at the University of Waterloo. Before that, she earned an MSc in Applied and Computational Mathematics from Simon Fraser University and a BSc in Statistics from Huazhong University of Science and Technology.



A Generative Model for Networks Grounded in Bayesian Modeling Principles

Federica Zoe Ricci, Swarthmore College, PA

Graphs play a crucial role in many applications, ranging from drug discovery to social network modeling and astronomy. We propose a generative model for graphs that matches or exceeds the state-of-the-art, but can scale to regimes where their training fails, and is orders of magnitude faster at generation. Our approach is inspired by the Aldous-Hoover theorem, a classic representation theorem that characterizes any probability distribution of graphs that is invariant to permutations of node indices (i.e., exchangeable). This theorem represents the foundation of many classic Bayesian models for networks, and establishes that three ingredients are needed: a graph-level variable, a set of node-specific variables, and a so-called graphon function that maps graph- and node-level variables to the probability of an edge between any node pair. Given a training set of graphs, our exchangeable Variational Graph AutoEncoder (xVGAE) learns an approximation to their underlying graphons, as well as graph- and node-level latent variables. We show that our xVGAE encodes interpretable representations of graph- and node-level properties, and can (quickly, in one shot) generate new graphs that closely match key training set statistics.

Bayesian Nonparametric Dependent Edge-Exchangeable Network Models for Criminal Data

Louise Alamichel, Bocconi University

Criminal networks are complex interaction systems characterized by sparsity, hidden structures, and incomplete information, posing significant challenges for statistical modeling. In this work, we propose a Bayesian nonparametric framework for modeling interaction data using dependent edge-exchangeable network models, built on the Product Centred Dirichlet Process (PCDP). To account for community structures commonly observed in criminal organizations, we further extend the model to a nested formulation, yielding a block edge-exchangeable model in the spirit of recent developments in node-level community detection within interaction processes. The proposed framework provides a principled and flexible approach for modeling bipartite criminal networks while incorporating dependence and hierarchical block structure.

Latent Network Models to Account for Noisy, Multiply Reported Social Network Data

Martina Contisciani, Center for Critical Computational Studies, Goethe University Frankfurt

Social network data are often constructed by incorporating reports from multiple individuals. However, it is not obvious how to reconcile discordant responses from individuals. There may be particular risks with multiply reported data if people's responses reflect normative expectations' such as an expectation of balanced, reciprocal relationships. In this talk, I will present a probabilistic model that incorporates ties reported by multiple individuals to estimate the unobserved network structure. In addition to estimating a parameter for each reporter that is

related to their tendency of over- or under-reporting relationships, the model explicitly incorporates a term for mutuality, the tendency to report ties in both directions involving the same alter. Our model's algorithmic implementation is based on variational inference, which makes it efficient and scalable to large systems. We apply our model to data from a Nicaraguan community collected with a roster-based design and 75 Indian villages collected with a name-generator design. We observe strong evidence of mutuality in both datasets, and find that this value varies by relationship type. Consequently, our model estimates networks with reciprocity values that are substantially different than those resulting from standard deterministic aggregation approaches, demonstrating the need to consider such issues when gathering, constructing, and analysing survey-based network data.

Bayesian Poisson-Randomized Gamma Tensor Factorization with Application to International Trade Flows

Jie Jian, University of Chicago

We study sparse semi-continuous tensor data with excess zeros, heavy right tails, and slice-specific dispersion. Such features arise naturally in monetary-valued multi-way data, such as international trade, where most exporter--importer--product--year cells are zero while positive values are continuous and highly variable. To model these data, we propose a Bayesian hierarchical tensor factorization model that places a low-rank CP structure on a latent Poisson rate tensor and couples it with a conditional Gamma model for positive outcomes, with rate parameters that can vary across slices within a mode. The model therefore separates the occurrence and magnitude of positive observations while borrowing strength across all tensor dimensions through a shared low-rank latent structure. To scale posterior inference to large arrays, we develop a hybrid variational--Monte Carlo algorithm that combines efficient coordinate ascent updates with a partially collapsed augmented-data sampler. Applied to approximately 60 million trade flows, the method surfaces multiway dependence across exporters, importers, products, and years that is difficult to recover from gravity-type or pairwise network analyses, which do not jointly model the product and temporal dimensions.



Session Info

S123

TECHNICAL STATISTICAL SESSION

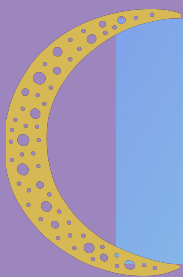
INVITED SESSION

October 6th, 22:00 - 23:00 UTC

Borrowing Strength: Statistical Strategies for Imperfect Data

USE SCHED TO ACCESS THE ZOOM ROOM

This session highlights how modern biostatistics tackles a shared challenge across very different problems: making valid, efficient inference when the data we have is incomplete, limited, imbalanced, or imperfectly comparable to the question we want to answer. Talks in this session explore how statisticians "borrow strength" – from historical or external data sources, from related but heterogeneous datasets, from carefully constructed causal designs, and from generative models that learn the structure of missingness itself – to push beyond the limitations of any single dataset. Together, these four talks, spanning clinical trials, large-scale survey methodology, high-dimensional prediction, and observational causal inference, showcase the breadth of ideas at the intersection of biostatistics, machine learning, and causal reasoning, and the methodological creativity required to turn imperfect data into trustworthy conclusions.

ORGANIZER: Xiaoting Chen, New York University**CHAIR:** Yajun Mei, New York University

Speaker Bios

S123

**MS. XIAOTING CHEN**

New York University

Xiaoting Chen is a 4th-year Biostatistics PhD candidate at New York University School of Global Public Health. Her research focuses on dynamic borrowing methods for clinical trial design, with an emphasis on frequentist frameworks that combine historical and concurrent control data while maintaining rigorous Type I error control.

**DR. YUYU(RUBY) CHEN**

New York University

Yuyu(Ruby) Chen recently graduated from NYU School of Global Public Health with PhD in Biostatistics. Her PhD work focused on developing and evaluating advanced missing data imputation frameworks tailored for healthcare research, which covers the aspects including clinical trials, multi-source data and national surveys.

**MS. JIANAN ZHU**

New York University

Jianan is a rising fifth-year PhD student in Biostatistics. Her main research interest is on design-based causal inference. She is interested in observational studies, randomized trials, sensitivity analysis, and their applications in health policy research, infectious disease research, and social sciences.

**MS. IRIS ZHANG**

New York University

Iris is a 4th-year Biostatistics PhD student at New York University School of Global Public Health. Her research focuses on statistical machine learning and high-dimensional inference, with an emphasis on transfer learning, representation learning, and AI-driven methods for healthcare. She is interested in building theoretically grounded methodologies that enable robust and efficient learning from complex, heterogeneous, and large-scale biomedical data.



Dynamic Borrowing of External Controls in Clinical Trials*Xiaoting Chen, New York University*

In randomized clinical trials, sample size constraints arising from practical and ethical considerations often motivate borrowing external historical information, such as prior randomized studies or Electronic Health Record (EHR) data, to augment the control arm. In the first part of this talk, we review the existing dynamic borrowing methods, including power priors and hierarchical modeling frameworks that adaptively incorporate external historical information. While these approaches can substantially improve estimation efficiency and power, they may induce Type I error inflation under prior-data conflict and are less well-developed for longitudinal data. In the second half, motivated by a recent Ritlecitinib trial for Alopecia Areata, where extrapolating unobserved longitudinal endpoints from a 24-week control to a 36-week treatment target was ethically necessary, we propose a novel dynamic borrowing framework for continuous longitudinal outcomes under a linear mixed-effects (LME) model with slope constraints. Hope our research can stimulate further methodological development on dynamic borrowing that achieves efficiency gains while maintaining the Type I error control in clinical trials.

TabSODA: Tabular Diffusion based Imputation with Skip Pattern Detection and Ordinal Awareness*Yuyu Chen, New York University*

Missing data imputation in large-scale surveys faces two challenges that are not well handled by current tabular diffusion methods. First, structural skips, cells made inapplicable by questionnaire design, should not be imputed but are often conflated with item nonresponse. Second, ordinal responses encode ordered categories, yet most pipelines treat them as nominal levels through one-hot or analog-bit encodings. We introduce TabSODA (Tabular diffusion with Skip pattern detection and Ordinal Awareness), an Expectation-Maximization (EM)-based diffusion imputer built on the Elucidated Diffusion Model (EDM) framework. TabSODA propagates structural skips through the denoising loss and reverse-time sampler, and represents ordinal variables with cumulative-probit scalar latents while retaining analog-bit encodings for nominal variables. When a codebook skip mask is available, TabSODA uses it directly; otherwise, the TabSODA+SKIP variant estimates the mask from raw responses and questionnaire order using a CART-based skip-pattern miner. On Population Assessment of Tobacco and Health (PATH) study and the National Survey on Drug Use and Health (NSDUH), two nationally representative U.S. surveys, TabSODA reduces ordinal MACE by up to 23.7%(relative) and improves categorical accuracy by up to 9% over the strongest baseline across MCAR, MAR, and MNAR masking. The skip miner achieves near-perfect precision on both datasets, allowing TabSODA+SKIP to closely track the codebook-mask variant.

Matching is a widely used causal inference design that aims to approximate a randomized experiment using observational data by forming matched sets of treated and control units based on similarities in their covariates. Ideally, treated units are exactly matched with controls on these covariates, enabling randomization-based inference for treatment effects as in a randomized experiment, under the assumption of no unobserved covariates. However, inexact matching often occurs, leading to residual covariate imbalance. Previous matched studies have typically overlooked this issue and relied on conventional randomization-based inference, as long as some covariate balance criteria are met. Recent research has shown that this approach can introduce significant bias and proposed methods to correct for bias arising from inexact matching in randomization-based inference. These methods, however, are primarily focused on the constant treatment effect and its extensions (i.e., Fisher's sharp null) and do not apply to average treatment effects (i.e., Neyman's weak null). To address this gap, we introduce a new method, inverse post-matching probability weighting, for conducting randomization-based inference for average treatment effects under inexact matching. Our theoretical and simulation results indicate that, compared to conventional randomization-based inference methods, our approach significantly reduces bias and improves coverage rates in the presence of inexact matching.

Advancements in Transfer Learning -- Leveraging Heterogeneous Data for Better Predictions*Iris Zhang, New York University*

Transfer learning has emerged as a powerful framework for improving prediction in settings where the target dataset is limited but related information is available from multiple external sources. A central challenge is that these auxiliary datasets often differ substantially from the target population in terms of covariate distributions, feature spaces, or underlying data-generating mechanisms. Effectively leveraging such heterogeneous data while avoiding negative transfer remains an active area of statistical research. We propose Faster (Feature Alignment and Structured Transfer via Efficient Regularization), a two-stage framework for high-dimensional generalized linear models with partially overlapping feature spaces, in the regime where the total feature dimension may grow much faster than any individual sample size. We establish a high-probability guarantee for sparse recovery in the alignment step and derive an error bound for the final estimator. Simulation studies and real-data analysis demonstrate the effectiveness of this method.

Randomization-Based Inference for Average Treatment Effects in Inexactly Matched Observational Studies.*Jianan Zhu, New York University*

Session Info

S124

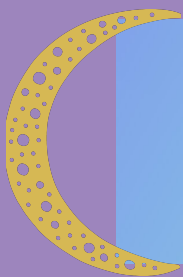
CLOSING SESSION

October 7th, 00:00 - 1:00 UTC

Closing Session

USE SCHED TO ACCESS THE ZOOM ROOM

In our closing session we will celebrate the 24 hours we spent together in this conference! We will hear from you about sessions that were impactful and engaging, your “ah ha” moments, and lessons learned from our speakers.

ORGANIZER: IDWSDS Program Committee**CHAIR:** Samantha Seals, University of West Florida

Session Info

S200

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 3:00 - 3:30 UTC

Absorbing Markov Chain Parameter Estimation Under Data Scarcity: A Comparative Study of Analytical and Monte Carlo Methods in Neonatal Care

USE SCHED TO ACCESS THE ZOOM ROOM

Neonatal mortality remains a major global public health challenge, with an estimated 6,200 newborns dying daily, mostly in settings where patient records are scarce. For data-scarce neonatal units, a key question arises: when transition data is limited, does it matter whether Analytical Estimation or Monte Carlo Simulation is used to model patient outcomes?

This study addresses that question using a neonatal dataset of 6,000 daily state transitions across 2,000 patients in Ghana. An absorbing Markov chain with four states: Hospital Admission, Neonatal Intensive Care Unit (NICU), Recovered, and Death was used. Both methods were evaluated across six data levels with 500 independent replications using bias, variance, standard deviation, and mean squared error.

Both methods perform similarly with adequate data and degrade equivalently under data scarcity because they share the same estimated transition matrix. Expected time to absorption is more sensitive to limited data than absorption probabilities, and 500 observed transitions emerge as the minimum reliable threshold. These findings provide an evidence-based data standard for resource-constrained healthcare systems.

Keywords: Absorbing Markov Chain, Monte Carlo Simulation, Analytical Estimation, Absorption Probability, Expected Time to Absorption, Sensitivity Analysis.

Authors: Emmanuella Frimpong, Dr. Irene Kafui Vorsah Amponsah, PhD (Visiting Lecturer at Ohio University)

ORGANIZER: Emmanuella Frimpong, Miami University

CHAIR: Irene Kafui Vorsah Amponsah, University of Cape Coast



Speaker Bios

S200



MS. EMMANUELLA FRIMPONG

Miami University

Ms. Emmanuella Frimpong is a graduate of the African Institute for Mathematical Sciences (AIMS) Ghana, where she completed her Master of Science in Mathematics, undertaking research on the topic “Comparing Monte Carlo Simulation and Analytical Estimation Methods for Absorbing Markov Chain Parameters Under Limited Transition Data”. She holds a funded position as a graduate teaching assistant in the Department of Statistics at Miami University in Oxford, Ohio, where she is pursuing a second Master of Science in Statistics with a focus on time series analysis and forecasting. Ms. Frimpong received her Bachelor of Science in Mathematics and Statistics from the University of Cape Coast, Ghana, in 2024, graduating as the Overall Best Student in the Department of Mathematics. She served as a teaching assistant in the department where she demonstrated her passion to advance statistical education and mentor the next generation of women in Statistics and Data Science.



Session Info

S201

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 00:30 - 1:00 UTC

Vision-Language Models and the Harms of Covert Sexualization

USE SCHED TO ACCESS THE ZOOM ROOM

Vision-language models (VLMs) increasingly mediate how bodies are described, moderated, and rendered in online spaces through content moderation, AI-generated images and videos, and descriptions of real bodies. Prior work establishes that VLMs sexually objectify bodies that are partially clothed more than fully clothed ones, and that plus-size bodies — particularly those belonging to women and AFAB persons — are disproportionately censored on social media platforms for being “inappropriate.” However, scholars lack the tools to detect when a model *objectifies* a body, rather than merely describing it, and whether this behavior differs across body *shapes*, rather than just body sizes. This talk describes a study in which we investigate these phenomena using swimwear try-on images matched across body morphology, comparing high-contrast silhouettes (i.e., a smaller waist relative to hips and bust) against lower-contrast silhouettes under identical prompting conditions. We ask whether VLMs describe these bodies differently despite equivalent context and whether unwanted objectification or descriptive drift disproportionately burdens those with high-contrast bodies, with a focus on covert (i.e., safety guardrail compliant) sexualization. By making the harms of these phenomena visible and measurable, this work gives researchers and auditors the ability to hold systems accountable when they objectify or censor women and AFAB persons based on their appearance.

ORGANIZER: Maimuna Majumder, Harvard Medical School & Boston Children's Hospital

CHAIR: Danielle Whisnant



Speaker Bios

S201

**PROF. MAIMUNA MAJUMDER**

Harvard Medical School & Boston Children's Hospital

Dr. Maimuna (Maia) Majumder (she/they), MPH, PhD (MIT '18) is an Assistant Professor and Inaugural Peter Szolovits Distinguished Scholar in the Computational Health Informatics Program at Harvard Medical School and Boston Children's Hospital. She is a computational epidemiologist specializing in the application of artificial intelligence and digital data — namely, search query, news & social media, and chatbot data — to societally relevant problems, including women and children's health.



Session Info

S202

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED PANEL

October 6th, 17:00 - 18:00 UTC

One Size Doesn't Fit All: Tailoring Mentorship Across Career Stages, Challenges, and Timelines

USE SCHED TO ACCESS THE ZOOM ROOM

Traditional mentorship models often assume a single, monolithic structure: an experienced senior professional guiding an early-career colleague over years. Effective mentorship is rarely universal. Successful developmental relationships require deliberate alignment with an individual's specific career stage, immediate professional hurdles, and appropriate engagement timelines. This presentation will explore strategies for cultivating responsive mentorship frameworks. From early-stage onboarding and skill acquisition to mid-career strategy and senior leadership shifts, the panel will consider how mentoring needs evolve across distinct transitions. We will discuss the distinct value of both short-term, task-focused, and longitudinal developmental partnerships. Participants will leave with practical frameworks for mentors and mentees to assess context, define clear mutual expectations, and build adaptable partnerships that deliver measurable impact.

ORGANIZER: Nairanjana (Jan) Dasgupta, Washington State University; Donna LaLonde, American Statistical Association

CHAIR: Umut Ozbek, Caucus for Women in Statistics and Data Science



Speaker Bios

S202

<http://www.math.wsu.edu/faculty/ndasgupta/>


NAIRANJANA (JAN) DASGUPTA
Washington State University

Nairanjana (Jan) Dasgupta is a Boeing Distinguished Professor of Science and Mathematics and Regents Professor and Chair Department of Mathematics and Statistics at Washington State University. She has been a fellow of American Statistical Association (ASA) since 2018. She has been part of the Advisory Board for Academic Data Science Alliance (ADSA) and currently is on the Board for International Indian Statistical Association. In 2026 she is serving as the President for WNAR-IBS. She is the past-President for Caucus for Women in Stats (CWS) and is the chair-elect of the JEDI outreach committee of the ASA. She was the past chair for the Committee for Women in Statistics and past president of Caucus for Women in Statistics and Data Science.



DONNA LALONDE
American Statistical Association

Donna LaLonde is the Associate Executive Director of the American Statistical Association. Her responsibilities include providing support to ensure progress on the ASA strategic initiatives. She supports the development and implementation of guidelines and reports endorsed by the ASA Board of Directors, supports presidential initiatives, and guides all professional development activities. Prior to joining the ASA, she was a faculty member in the Department of Mathematics and Statistics at Washburn University. In addition to a faculty role LaLonde served as Associate Dean, College of Arts and Sciences, Associate Vice President for Academic Affairs, Director of Assessment, and Interim Chair, Department of Education.



DR. SHARI MESSINGER
University of Miami Miller School of Medicine

Shari Messinger, Ph.D., FACTS, is Professor of Biostatistics and Director of the Division of Biostatistics and Bioinformatics in the Department of Public Health Sciences at the University of Miami Miller School of Medicine. She also directs the Biostatistics, Epidemiology and Research Design Program of the Miami Clinical and Translational Science Institute and the Biostatistics Collaboration and Consulting Core.

Dr. Messinger chairs the American Statistical Association's Committee on Women in Statistics, serves as President of the ASA South Florida Chapter and on the ASA Scientific and Public Affairs Advisory Committee, and sits on the Board of Directors of the Association for Clinical and Translational Science, which elected her an inaugural Fellow in 2026.



KAREN NIELSON
Georgia State University

Karen Emily Nielsen, PhD, is an Associate Professor of Biostatistics in the School of Public Health at Georgia State University whose research centers on statistical methodology for wearable technologies, time-intensive longitudinal data, and data harmonization. She develops and translates methods that help researchers integrate complex behavioral, physiological, and sensor-derived data, with applications spanning public health, clinical research, and global health. Her work has contributed to methodological advances in mobile and wearable data analysis, informed national public health surveillance efforts, and supported diverse interdisciplinary research teams.



Session Info

S203

TECHNICAL STATISTICAL SESSION

OTHER

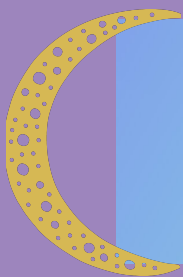
CONTRIBUTED SESSION

October 6th, 4:00 - 5:00 UTC

AI and Data Science 1

USE SCHED TO ACCESS THE ZOOM ROOM

Who Gets Heard? Investigating Accent and Gender Disparities in Automatic Speech Recognition | Bridging Data Silos: A Two-Step Unsupervised Approach to Record Linkage | Mining the Foundations of Truth: Extracting Structured Knowledge Graphs from Unstructured Classical Texts

ORGANIZER: IDWSDS Program Committee**CHAIR:** Su Na Chin, Universiti Malaysia Sabah

Speaker Bios

S203



MS. TISHA PRASAD
STEM4Change

Tisha Prasad is the co-founder and co-president of STEM4Change, a nonprofit organization dedicated to expanding access to STEM education by teaching students foundational concepts like coding. She also serves as Head of Web Development at VolunteerConnect, where she leads the development of technology that connects volunteers with nonprofit organizations and service opportunities. As a Women in Data Science ambassador, she is leading the development of Lift Our Voices, a project focused on closing the gender data gap to improve the accuracy and inclusivity of automatic speech recognition across diverse vocal patterns. Her interests lie at the intersection of technology, education, and social impact, with a focus on developing equitable technologies that empower communities and broaden opportunities for learning.



MRS. ZATURRAWIAH ALI OMAR
Australian BioCommons

Zaturrawiah Ali Omar is a lecturer at Universiti Malaysia Sabah (UMS) and a PhD candidate at Universiti Kebangsaan Malaysia (UKM), where her research focuses on record linkage, machine learning, and road traffic injury surveillance in Malaysia. Her work centres on using the Unsupervised Random Forest (URF) algorithm to link and match records without relying on labelled training data.



MS. NORHAFIZA HAMZAH
Universiti Malaysia Sabah

Norhafiza is a lecturer and researcher at Universiti Malaysia Sabah, developing knowledge graph and machine learning frameworks for fact-checking religious misinformation on social media. She teaches courses in Artificial Intelligence, Machine Learning, and Computer Science courses at the Faculty of Science and Technology, UMS, bringing both applied research and classroom experience to her work. Her broader research interests include machine learning, data analytics and data mining, with a focus on building intelligent systems that make sense of complex real-world data.



Who Gets Heard? Investigating Accent and Gender Disparities in Automatic Speech Recognition*Tisha Prasad, STEM4Change*

Automatic speech recognition (ASR) systems have achieved remarkable performance in recent years, yet concerns remain regarding their robustness across speakers with diverse linguistic backgrounds and gender identities. Differences in pronunciation caused by non-native accents, alongside variation in speech characteristics across gender, can lead to unequal recognition performance and introduce potential fairness and quality-of-service issues. This study evaluates the Whisper-tiny.en ASR model using the GMU Speech Accent Archive, measuring ASR performance across native and non-native English speakers from over 150 accent groups, with a focus on how gendered performance differences manifest across accents. Performance differences are quantified using Word Error Rate, Character Error Rate, Match Error Rate, and Word Information Lost. The results reveal substantial disparities across accents, with East and Southeast Asian language backgrounds (e.g., Mandarin, Cantonese, Japanese, Korean, Vietnamese, Lao), as well as a subset of African language backgrounds (e.g., Somali, Bai, Miskito, Fang) exhibiting markedly higher error rates than speakers with English and European language backgrounds, (e.g. Afrikaans, Dutch, German, Swedish, Hungarian). Across all accents, female speakers achieve slightly lower error rates than male speakers, suggesting that while accent is the primary source of performance variation, gender-related differences persist.

fundamentally as a complex data mining problem. Using a corpus of Islamic jurisprudence (Hadiths) as a primary case study, we explore the methodological challenges of mining unstructured, classical texts to discover and extract structured data representations. The presentation will detail the end-to-end data mining pipeline, including NLP-driven entity recognition, coreference resolution, and relationship extraction, required to transform raw text into a mathematically traversable Knowledge Graph. By demonstrating how to mine and structure these deeply embedded contextual rules, this talk illustrates how rigorous data mining serves as the critical first step in enabling accurate, culturally resilient automated verification systems on a global scale.

Bridging Data Silos: A Two-Step Unsupervised Approach to Record Linkage*Zaturrawiah Ali Omar, Universiti Malaysia Sabah*

Linking records across databases, whether to identify duplicates or match the same entity across sources, is a foundational step in producing trustworthy, connected data. The most established approach, probabilistic record linkage, still depends on carefully set similarity thresholds, a process that becomes difficult without access to labelled "gold standard" data to validate against. Supervised machine learning approaches can offer an alternative, but they introduce their own dependency on labelled training data, which is often expensive or simply unavailable. This talk presents a two step approach built on the Unsupervised Random Forest (URF) algorithm, which generates a similarity score for record pairs without requiring any labels at all. We explore how this single algorithm can serve two purposes: detecting duplicate records directly, and automatically selecting high quality training data to support supervised classification, offering a path toward record linkage that is less reliant on labelled data.

Mining the Foundations of Truth: Extracting Structured Knowledge Graphs from Unstructured Classical Texts*Norhafiza Hamzah, UMS*

As the volume of digital information grows, global artificial intelligence systems increasingly struggle to process highly specialized, culturally nuanced data without hallucinating. Aligning with the IDWSDS 2026 theme of "Creating Global Connections", this talk frames domain-specific fact-checking



Session Info

S204

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

PLENARY TALK

October 6th, 7:00 - 7:30 UTC

Creating Global Connections Through Early Statistical Leadership in Rare Disease Trials

USE SCHED TO ACCESS THE ZOOM ROOM

In complex clinical research, collaboration begins well before database lock or final analysis. It begins when statisticians are involved early enough in protocol development to shape what is measurable, sustainable, and meaningful for patients. This presentation uses a rare pediatric dermatology trial as a case study to show how early statistical input can prevent avoidable missing data before the first participant is enrolled.

The original design was scientifically ambitious but operationally burdensome, with frequent in-clinic visits, narrow assessment windows, and substantial caregiver burden. Statistically, this created a foreseeable risk of informative missingness, loss to follow-up, and reduced interpretability in a small and vulnerable population. Early statistical leadership helped reframe the design around feasibility as well as rigor.

The talk will illustrate how collaboration among statisticians, clinicians, programmers, operations teams, and digital partners can reduce this risk. Examples include identifying burden-sensitive endpoints, anticipating missing-data pathways, and considering remote ePRO, image capture, or AI-supported skin assessment to replace selected in-person visits without compromising data quality.

The broader message is that statistics is not only an analysis discipline; it is also a design and connecting discipline.

ORGANIZER: Shrutimita Pokhariyal, PHASTAR**CHAIR:** Shrutimita Pokhariyal, PHASTAR

Session Info

S205

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 8:00 - 9:00 UTC

Time-to-Event Data in Biometry: Challenges, Biases and Perspectives

USE SCHED TO ACCESS THE ZOOM ROOM

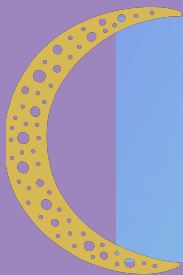
In applied medical statistics, researchers encounter a wide variety of data structures. Among these, time-to-event data play a central role including survival times, transitions between disease stages, and sequences of medical events or progressions of patients' states over time.

A broad spectrum of analytical approaches is available, ranging from (semi-)parametric methods such as the Cox proportional hazards model to flexible spline-based and other advanced modeling strategies. Depending on the motivation we encounter unique problems as well as unique solutions on which we want to shed light on in this session. Each setting introduces its own methodological challenges and potential sources of bias.

We aim to highlight key issues that arise when working with time-to-event data, including common pitfalls, structural complexities, and opportunities for methodological innovation. We will present different perspectives on model choice, interpretation, and discuss how careful consideration of the underlying data-generating mechanisms can inform both statistical analysis and the design of clinically relevant studies.

ORGANIZER: Marilena Müller, German Cancer Research Center, Heidelberg, Germany

CHAIR: Marilena Müller, German Cancer Research Center, Heidelberg, Germany



Speaker Bios

S205

**MISS JUDITH VILSMEIER**

Institute of Statistics, Ulm University

Judith Vilsmeier is a doctoral candidate in Biostatistics at Ulm University and a research assistant at the Institute of Statistics. She received both her B.Sc. and M.Sc. in Mathematical Biometry from Ulm University. Her current research interests involve nonstandard event histories and estimation of complex outcomes in non-Markov multi-state models. In addition to her research, she contributes to teaching in the areas of survival analysis, event history analysis, and advanced statistics.

**DR. SANDRA SCHMELLER**

Institute of Statistics, Ulm University

Sandra Schmeller is a postdoctoral researcher in Biostatistics at the Institute of Statistics, Ulm University. She received both her B.Sc. and M.Sc. in Mathematical Biometry from Ulm University and completed her PhD in Biostatistics at Ulm University. Prior to her doctoral studies, she worked as a statistician at the German national registry for stem cell transplantation and cell therapy. Her research focuses on methodological developments and applications of multistate models in medical research, particularly in the context of hematopoietic stem cell transplantation. In addition to her research activities, she contributes to teaching in statistics and biostatistics at Ulm University.

**DR. MARILENA MÜLLER**

German Cancer Research Center

Marilena Müller is a postdoctoral scientist in the Biostatistics department at the German Cancer Research Center in Heidelberg, Germany. She received her B.Sc., M.Sc. and Dr. rer. nat. at the Mathematical Institute at Heidelberg University. The field of studies concerning her dissertation include Rebolledo-type limit theorems applied to Hawes processes. Additional research fields include time-to-event methods, mixed model approaches and stochastic processes. Application to medical data as well as teaching are additional contributions to the research world.

**MISS EMA POŽEK**

University of Ljubljana

Ema Požek is a PhD student in statistics at the Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana. Her methodological interests lie in survival analysis and simulation studies, and her research focuses on developing statistical methods for lead time and overdiagnosis estimation in cancer screening programmes. She previously worked as a biostatistician at the German Cancer Research Center (DKFZ), where she was involved in the analysis of clinical trials. Ema holds a master's degree in mathematical statistics from the University of Zagreb (2022) and is actively involved in collaborations with clinicians and researchers on applied projects in medical research and statistics.

The challenge of time-to-event analysis: Specialties in a competing risk setting

Judith Vilsmeier, Institute of Statistics, Ulm University, Germany

Time to first event analysis often considers a combined endpoint like time from a hematopoietic stem cell transplantation to relapse or death whatever comes first, or time until hospital admission or death. Although, a relapse or death after transplantation in the first example means a failure of the transplantation, it is often important to distinguish the type of events because a relapse allows further treatment. The talk will explain the extension to competing risks where the type of event is considered. This setting provides more information about the behavior of the disease and possible covariables. A focus will be on the difference of hazard and probability estimation and its interpretation. Nonparametric and semi-parametric models and their assumptions will be explained. A data example from leukemia patients shows that results on hazard level are needed to explain results for the probability that seem implausible at first glance. Furthermore, we suggest to estimate a time dependent effect piecewise constant to facilitate the interpretation. This talk is suitable for those who haven't had much experience with competing risks.

The challenge of time-to-event analysis: New insights when extending the model

Sandra Schmeller, Institute of Statistics, Ulm University, Germany

Clinical trials often compare a treatment to a control group with respect to multiple possible combined time-to-event endpoints like time to hospitalization and time to death. Thereby, the first endpoint may occur more than once (recurrent), whereas the second endpoint is absorbing, that is, after observing the second endpoint an individual can no longer experience any other events. However, usually only the time until the first occurrence of an event for a patient is analyzed. Inclusion of all observed events in the analysis can increase the power and provides a more complete picture of the disease but needs more sophisticated methodology. The talk will give a stepwise guidance on how to extend the time-to-first event model to complex multistate methodology, where multiple events are incorporated. We thereby consider nonparametric methodology and semi-parametric methods and show how they are related. Special attention is given on the prerequisites of the models, e.g., the Markov property, and their (causal) interpretation. A data example from the Interdisciplinary Network for Heart Failure where the efficacy of a nurse-coordinated disease management program is compared to usual care for patients that were first hospitalized for systolic heart failure will demonstrate theoretical results and show that more insights are gained when considering multiple events. This talk is an extension of the previous talk on competing risks.

Estimating a treatment effect for survival when assignment to the target population is partially hidden using restricted mean survival times

Marilena Müller, Biostatistics, German Cancer Research Center,

Heidelberg, Germany

We consider a two-arm randomized clinical trial with time-to-event endpoint. Patients in the control arm solely receive standard of care (SOC) treatment whereas patients in the experimental arm are offered personalized treatment. However, a big portion of patients in the experimental arm will not receive personalized treatment due to various causes; the main causes are that no personalized treatment is available and that patients do not consent to treatment. These patients also receive SOC.

The estimand of interest is the treatment effect in the population that "potentially receive personalized treatment", i.e. that would receive personalized treatment if randomized to the experimental arm. Direct estimation of this effect is not feasible, since information on which patients in the SOC arm would have received personalized treatment if randomized to the experimental arm is missing.

In the first step of this approach we have to estimate the survival function of the subpopulation of the treatment arm nonparametrically. For the SOC arm the usage of the Kaplan-Meier-estimator is possible. These estimated survival curves are then compared using the restricted mean survival time and hypothesis testing is performed via bootstrap resampling.

This work presents a novel rigorous model for the analysis of treatments effects in the presence of mixtures or asymmetric trials.

Survival analysis of cancer screening: biases and lead time estimation

Emma Požek, Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia

Cancer screening programmes aim to detect disease at an early stage, when treatment options may be broader and chances of recovery may be greater. However, evaluating the benefits of screening programmes is not straightforward, as naïve comparisons of survival of those invited and those not invited to the programme are biased. Apparent improvements in survival among screen-detected cases may simply reflect earlier diagnosis (lead time bias), preferential detection of slower-growing disease (length time bias), or detection of cases that would never have been detected in the absence of screening (overdetection). These mechanisms can inflate observed survival without necessarily reducing cancer-specific mortality. In this talk, I will first give a concise overview of the main biases that affect observational evaluation of screening. I will then focus on recent work by Vratnar and Pohar Perme (2023), who formalized the relevant target estimand for programme evaluation and showed that the total bias in survival can be decomposed into lead time, length time, and overdetection components. Because lead time plays a central role in this framework, its estimation is an important methodological issue. I will briefly review commonly used approaches to lead time estimation, with emphasis on their assumptions, strengths, and limitations.



Session Info

S206

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 9:00 - 9:30 UTC

Beyond Simulation: Benchmarking for Method Comparison in Statistics and Machine Learning

USE SCHED TO ACCESS THE ZOOM ROOM

When initiating a statistical or machine learning analysis, one of the first and most consequential questions is: which method should be used for a given dataset? Traditional approaches for comparing methods include theoretical derivations and data simulation studies, which provide insight into method performance under controlled conditions. However, these approaches may not fully reflect the complexity of real-world data. Benchmarking—systematic comparison of methods across many real datasets—offers a complementary approach that can improve generalizability and provide practical guidance. Despite its common use in computer science, benchmarking remains underutilized in statistical methodology as new methods continue to emerge.

In this talk, I will discuss benchmarking in the context of statistical and machine learning research and contrast it with theory and simulation. I will outline key principles for conducting rigorous benchmarking studies and illustrate them using two case studies: benchmarking random forest variable selection methods for categorical and continuous outcomes and comparing methods for time-to-event data using the mlr3 framework. Together, these examples demonstrate how benchmarking can enhance scientific rigor, provide practical guidance for method selection, and support more transparent and reproducible methodological research.

ORGANIZER: Jaime Speiser, Wake Forest University School of Medicine

CHAIR: Yi-Ju Lee, Institute of Statistical Science, Academia Sinica



Speaker Bios

S206

**DR. JAIME SPEISER**

Wake Forest University School of Medicine

<https://sites.google.com/view/jaimespeiser/home>

Dr. Speiser is a biostatistician focused on prediction modeling with applications in medicine. Her work involves developing novel machine learning methodology for prediction modeling, providing guidance on best practices for developing prediction models, and collaborating with medical researchers to develop models that inform clinical decision-making in the learning health system. Dr. Speiser received her B.S. in mathematics from Elon University and her M.S. in statistics from The Ohio State University. She completed her Ph.D. in biostatistics at the Medical University of South Carolina, where she was awarded the Distinguished Graduate Student Award at the top of her class. She serves as Co-Lead of the Biostatistics and Research Informatics Core of Wake Forest's Claude Pepper Older Americans Independence Center. Dr. Speiser holds leadership roles in the ASA's Statistics and Data Science in Aging Interest Group and the Section on Statistical Learning and Data Science.



Session Info

S207

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 9:30 - 10:00 UTC

On Similarity-Based Models for Bayesian Disease Mapping

USE SCHED TO ACCESS THE ZOOM ROOM

In the 1970s, the conditionally formulated Gaussian Markov random field (GMRF), known as the conditional autoregressive (CAR) model, was introduced in line with Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things." The CAR model uses W , the well-known adjacency matrix that encodes the neighbourhood structure of a spatial lattice (e.g., two areas are neighbours if they share a common border).

Since its introduction, the CAR model has undergone many adaptations, with numerous adaptive models proposed in the literature. However, almost all of these models (to the best of our knowledge, all except ours) still adhere to Tobler's law. Yet, much of the data collected today – often aggregated at the areal level for confidentiality or other reasons – does not necessarily follow this law.

In this talk, we will show that any extra information representing causes of or correlated with the phenomenon of interest can be used to define a similarity structure, rather than relying solely on geographical neighbourhoods. Using simulated data, we illustrate that similarity-based structures can be more effective than traditional neighbourhood-based structures for smoothing both local and global risks. We will show that the correct identification of high- and low-risk areas, crucial for public health planning and resource allocation, is better achieved when the similarity-based structure is used.

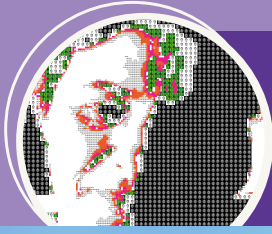
ORGANIZER: Helena Baptista, Management Information Centre (MagIC), NOVA Information Management School (NOVA IMS)

CHAIR: Yi-Ju Lee, Institute of Statistical Science, Academia Sinica



Speaker Bios

S207



DR. HELENA BAPTISTA

Management Information Centre (MagIC), NOVA

Helena Baptista is a highly experienced statistician, researcher, and educator, specializing in applied statistics, forecasting, and time series analysis. She has over 25 years of experience in the pharmaceutical industry, finance, and academia, with a strong background in statistical modeling, data analytics, and methodological research.

She holds a PhD in Statistics and Econometrics and has been actively involved in organizing and teaching courses in applied, basic, and advanced statistics, both in traditional and e-learning formats. As a Professor of Practice at NOVA IMS, she has led multiple courses on Time Series Analysis, Forecasting Methods, and Statistics for Enterprise Data Analytics, adapting content for virtual learning environments during the COVID-19 pandemic.

Her main research topic is on Bayesian disease mapping models, which will be the topic of the session.



TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 10:00 - 10:30 UTC

Optimal Model selection for incidence of Birth Asphyxia: NICU Centers in Greater Accra Region

USE SCHED TO ACCESS THE ZOOM ROOM

Birth asphyxia remains a major contributor to neonatal morbidity and mortality in low- and middle-income countries, particularly in sub-Saharan Africa. This study investigated the determinants of birth asphyxia among newborns using a Quasi-Poisson regression model to account for overdispersion in the count data. Secondary data comprising neonatal and maternal records were analysed using descriptive statistics, correlation analysis, and inferential modelling. The Quasi-Poisson model selected after diagnostic assessment confirmed overdispersion in the response variable, making it more appropriate than the standard Poisson model. The Quasi-Poisson regression model for predicting birth asphyxia is given as : $\text{Birth Asphyxia} = 1.742 + 0.385(\text{Birth Weight}) + 0.012(\text{Mothers Age}) - 0.088(\text{Gestational Age}) + 0.0401(\text{Sex}) + 0.301(\text{Mode of Delivery}) + 0.158(\text{Parity}) + 0.067(\text{SURVIVE})$

The results showed that birth weight, gestational age, mode of delivery, maternal age, parity, and fetal presentation were significant predictors of birth asphyxia. Specifically, lower birth weight and shorter gestational age were associated with a higher incidence of birth asphyxia, while caesarean delivery and abnormal fetal presentation increased the likelihood of adverse birth outcomes.

Keywords: Birth asphyxia, Quasi-Poisson regression, Gestational age, Neonatal outcomes, Maternal health, Ghana.

Authors:

1. Selina Dadzie (Mphil)
2. Irene Kafui Vorsah Amponsah (PhD)

ORGANIZER: Selina Dadzie, Student

CHAIR: Gabriella Tarantini, IRCCS Azienda Ospedaliero-Universitaria di Bologna



Speaker Bios

S208



MISS SELINA DADZIE
Student

Miss Selina Dadzie is a graduate student in Statistics at the University of Cape Coast, Ghana, where she is pursuing a Master of Philosophy (MPhil) in Statistics. Her research focuses on “Optimal Model Selection for Incidence of Birth Asphyxia: NICU Centers in Accra”, with particular emphasis on count data regression models and neonatal health outcomes.

She holds a Bachelor of Science in Statistics from the University of Cape Coast and has developed strong research interests in biostatistics, epidemiology, generalized linear models, and health data analytics. Her current work evaluates the performance of Poisson and Quasi-Poisson regression models in modelling neonatal mortality, with the aim of improving evidence-based decision-making in healthcare.

Beyond her research, Miss Dadzie is passionate about applying statistical methods to solve real-world public health challenges.



Session Info

S209

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 10:30 - 11:00 UTC

Temporal and Regional Variations of Effects of Daily Temperature on Annual Precipitation: A Functional Mixed Effect Model Approach

USE SCHED TO ACCESS THE ZOOM ROOM

Given the increasing threat of global warming, it is important to understand not only global weather patterns but also regional variations throughout countries. Functional Linear Mixed-effects Model (FLMM), an emerging statistical tool, provides a comprehensive framework for analyzing functional data (e.g., data viewed as a function or curve) with repeated observations, allowing researchers to identify patterns and relationships in the data. This study applies the functional linear mixed-effects model (FLMM) to recognize the effects of temporal and regional variations of short-time weather projection (monthly precipitation on daily temperature). We deployed FLMM using the daily temperature and monthly precipitation of nine weather stations (Dhaka, Chattogram (Patenga), Chattogram (Ambagan), Rajshahi, Khulna, Barisal, Sylhet, Rangpur and Mymensingh) of Bangladesh where each station shares the common population effects with their individual scalar covariate effects along with same slope functions. To estimate variance parameters and fixed-effects and random effects, the REML-based EM algorithm proposed by has been applied. Empirical results show significant differences in the effects of monthly precipitation on daily temperature among regions of Bangladesh. We anticipated that FLMM is an emerging model that can reveal the effects of monthly precipitation and the pace of daily temperature fluctuations over time.

ORGANIZER: Munniara Yesmin Munni, Jahangirnagar University

CHAIR: Gabriella Tarantini, IRCCS Azienda Ospedaliero-Universitaria di Bologna



Speaker Bios

S209



MS. MUNNIARA YESMIN MUNNI

Jahangirnagar University

Munniara Yesmin Munni is a statistician and Monitoring, Evaluation, and Learning (MEL) professional with a Master's degree in Statistics from Jahangirnagar University, Bangladesh. She currently works with the Christian Commission for Development in Bangladesh (CCDB). Her research interests center on public health, particularly health equity, healthcare disparities, and the impacts of climate change on health. With strong expertise in statistical analysis and the integration of quantitative and qualitative methods, she is dedicated to advancing evidence-based solutions that promote community health and equity and contribute to sustainable development goals to improve the well-being of all.



Session Info

S210

TECHNICAL STATISTICAL SESSION

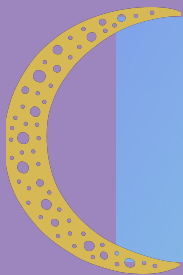
INVITED SESSION

October 6th, 12:00 - 13:00 UTC

Digitalisation of Public Institutions in Low Resource Settings

USE SCHED TO ACCESS THE ZOOM ROOM

Digitalizing public institutions in low-resource settings requires balancing technological innovation with systemic equity. Across Africa, infrastructure bottlenecks and high deployment costs often stall digital transformation. This session examines how public sector agencies can successfully transition from fragmented, legacy offline systems to unified digital platforms. We will discuss scalable hybrid architectures that maintain crucial offline functionality for underserved communities while enforcing centralised online data validation, optimizing institutional workflows, and improving financial reconciliation.

ORGANIZER: Chipso Zidana, NUST**CHAIR:** Chipso Zidana, NUST

Speaker Bios

S210



DR. LOUISA MUPARURI
ZIMSEC

Dr. Louisa Muparuri is an AI researcher, ICT executive, and digital transformation leader with expertise in artificial intelligence, data science, mathematical modelling, computing, examination systems, and public-sector innovation. She holds a PhD in Mathematics, an MSc in Operations Research and Statistics, an MBA, a BSc Honours degree in Computer Science, an Honours qualification in Data Science and Machine Learning and several industry professional certifications from Microsoft, Oracle and Google. She serves as Director of ICT and Examination Systems Development at the Zimbabwe School Examinations Council, where she leads the modernisation of national assessment platforms. Her research focuses on AI-assisted assessment, trustworthy AI, and scalable digital solutions for low-resource education systems in Africa.



MS. PHYLISS SHAMISO MAVEDZENGE
Bindura University of Science and Technology

Ms Phyliss Mavedzenge. Currently a Resident Innovator and Researcher at Zimbabwe Centre for High Performance Computing and a Business Analytics and Predictive Modelling, short course facilitator at Bindura University. Working on developing systems and models that cater to Zimbabweans and specializing mostly in Natural language processing for low resource languages like Shona. I have contributed to the development of the current police Automated Biometric Identification system and the current system being used by the Tertiary Education Service Council among other systems and I have a degree in Data Science and Informatics from the University of Zimbabwe and I graduated in 2025 and also presented at the African Women in Data Science 2025 virtual conference.



SALOME MAHEYA
African Energy Commission of the African Union

Salome Maheya, Senior Policy Officer for Energy Statistics in the African Energy Commission (AFREC) of the African Union. She is responsible for coordinating the production of annual Energy Balances, Energy Modelling and producing energy forecasts for Africa. She is responsible for training and assisting member states to establish their National Energy Information Systems which feed into the Africa Energy Information system. She coordinated the development of the Strategy for Harmonisation of Energy Statistics and is assisting in the development of other Energy policies and strategies. Salome provides guidance and support to African member states energy statistics focal persons on energy balance, energy efficiency indicators for end users, Sustainable Development Goals, CO2 Emissions and Nationally Determined Contributions. Salome holds a Masters Degree in Biostatistics and is a certified Energy Statistician and has more than 14 years' experience in energy data collection.



MS. LERDINIA VARAIDZO MAPEPA
Women's University in Africa

Ms. L. V. Mapepa is an Academic in the Faculty of Management, Department of Information Systems at the Women's University in Africa where she has served the institution for over five years across the various fields of teaching, learning and Innovation, Research and STEM initiatives. Ms. Mapepa is a self-driven Academic with a passion for excellence precisely in the broad field of Tech and Machine learning and Artificial Intelligence. She is a holder of various qualifications in IT including a Diploma and Advanced Diploma in IT Systems support, BSc Honors Degree in Information Systems, concluding an MPhil in Data Science and Machine learning and currently pursuing a Masters in Business Administration. She also takes pride in providing excellent service to all stakeholders while making positive impact in people's lives and in business especially women and girls.



From CDs to Connected Centres: Hybrid Online Candidate Registration for Exam Centres

Louisa Muparuri, ZIMSEC

National examination agencies often rely on legacy, offline desktop registration systems, which at the Zimbabwe School Examinations Council (ZIMSEC) created high operational risks and costs across over 11,000 centres. Physical data submission via removable media caused regional congestion and delayed bank reconciliations, leading to late underpayment detection and increased bad-debtor volumes. To address this, ZIMSEC implemented a hybrid candidate registration platform that supports offline capture for low-connectivity communities while requiring centres to submit a validated "final registration" online. This approach improves national visibility and accelerates near real-time reconciliation without excluding underserved areas. The paper describes the platform's architecture and pilot plan—covering centre onboarding, role-based access, and integration with payment workflows—and notes promising early results, including reduced congestion and faster discrepancy detection. It concludes with a practical evaluation framework focusing on reduced costs, faster cycles, lower debtor rates, and equitable participation.

Contrastive and Transfer Learning for Enhancing Contextual Representations in Shona News Classification

Phylliss Shamiso Mavedzenge, Bindura University of Science and Technology

LLMs have revolutionized NLP, but fine-tuning them for niche, domain-specific projects remains challenging—especially for low-resource languages with sparse digital archives. In Zimbabwe, Shona is spoken by over 10 million people, yet it lacks the massive electronic corpus required to train highly accurate models, causing it to be left behind by mainstream NLP developments.

This research addresses this gap by introducing a novel fine-tuning approach using a dynamic margin-based triplet loss function with a lambda parameter. Instead of relying on vast datasets, this method focuses on understanding underlying contextual relationships to effectively separate articles into distinct classes.

Using the MasakhaNEWS dataset (369 articles across four categories) and transfer learning, the proposed model generated embeddings for a separate dataset, which were then fed into a classification algorithm. Compared to a heavily trained base model evaluated on the same dataset, this context-focused approach improved classification accuracy by 14%, alongside improvements in other key metrics.

The findings demonstrate that massive training datasets are not strictly necessary for successful fine-tuning in low-resource settings. Instead, optimization techniques that prioritize deep comprehension of underlying contextual relationships offer a highly effective alternative when large scale electronic archives are unavailable

Transforming Official Energy Statistics through Data Science, Artificial Intelligence and Digital Innovation

Salome Maheya, African Energy Commission (AFREC), African Union

The growing demand for policy-relevant official statistics on sustainable development, energy transitions, and climate action is severely hindered in low-resource settings by fragmented, paper-based, and manual reporting systems. This presentation details how digital transformation modernizes public institutions by strengthening the entire statistical lifecycle—from collection to policymaking. Drawing on the African Energy Commission's (AFREC) collaboration with African Union Member States, it showcases practical digital solutions tailored for resource-constrained environments, including electronic data collection, integrated energy data warehouses, automated quality assurance, interactive business intelligence dashboards, and AI-driven data extraction. Using energy statistics as a case study, we illustrate how harmonized systems accelerate national planning, SDG 7 monitoring, and AU Agenda 2063 implementation. Ultimately, this abstract demonstrates that digitalization is a strategic investment in governance, highlighting the critical roles of data governance, institutional collaboration, and change management, with special emphasis on the rising leadership of women statisticians and data scientists driving public service delivery across Africa.

From Manual to Measurable "Digitalisation of Public Institutions in Low-Resource Settings

Lerdinia Mapepa, Women's University in Africa

Living in an era where data has become one of the world's most valuable resources practically Data in now referred to as the "The New Gold". Governments are no longer judged solely by the policies they formulate but increasingly by how effectively they use data and digital technologies to improve the lives of their citizens. The ability to deliver faster, smarter, and more transparent public services has become a defining characteristic of modern governance. The digitalization of public institutions has become central to improving governance, public service delivery, and evidence-based decision-making. While numerous studies have examined digital transformation from policy and technological perspectives, limited empirical research has investigated how artificial intelligence (AI) machine learning (ML) and predictive analytics can enhance digital decision-making within resource-constrained public institutions.

Digitalization should therefore not be viewed merely as replacing paper with computers or digitizing existing processes. True digital transformation involves fundamentally rethinking how institutions operate, make decisions, and engage with citizens. It is about creating intelligent public institutions that use data to anticipate problems, allocate resources effectively, and deliver services that are responsive, transparent, and inclusive.



Session Info

S211

TECHNICAL STATISTICAL SESSION

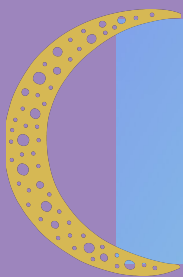
INVITED SESSION

October 6th, 13:00 - 14:00 UTC

Statistical Applications Across Various Domains From India

USE SCHED TO ACCESS THE ZOOM ROOM

Statistical applications and data analytics drive critical advancements across major global sectors. By leveraging historical data and predictive modeling, these methodologies improve efficiency, reduce risks, and optimize resource allocation worldwide. This session will focus on how statistical tools and methodologies enable data-driven decision making across different domains such as banking, business analytics, agriculture and healthcare in India. We have speakers from academia, government and different industries providing us with an inside journey of their work.

ORGANIZER: Sayantee Jana, Indian Institute of Technology**CHAIR:** Sayantee Jana, Indian Institute of Technology

Speaker Bios



DR. SUMAIYA SANDE
Wells Fargo

Dr. Sumaiya Sande is a Data Scientist and Credit Risk modelling professional with a PhD in Statistics and Applied Probability from the National University of Singapore. She specializes in machine learning, deep learning, statistical modeling, and credit risk analytics, with experience spanning healthcare, finance, and computer vision. Her industry experience includes roles at Wells Fargo, Northern Trust and a startup Qzense Labs, where she developed predictive models and scalable AI solutions. Sumaiya has published research in leading international journals and is passionate about translating advanced analytics into real-world business impact. Alongside her industry work, she has taught statistics and data science at the university level and actively shares knowledge through technical writing and mentoring. She is passionate about solving complex problems using data-driven approaches while continuously learning and innovating.



MS. SWASTIKA MOHAPATRA
Google

Swastika Mohapatra is a Research Data Scientist on the Analytics, Insights, and Measurement team within Google Ads. Over the past two years, her work has focused on building models to accurately measure digital ad campaign performance, sitting at the intersection of predictive modeling, causal inference, and privacy-centric measurement. She holds a Master's degree in Data Science from the Chennai Mathematical Institute and a Bachelor's degree in Statistics from Delhi University. Swastika is passionate about applying rigorous statistical and mathematical frameworks to solve high-velocity, real-world challenges in the tech industry.



DR. YOGITA GHARDE
ICAR-Directorate of Weed Research (ICAR-DWR)

Dr. Yogita Gharde is currently serving as Senior Scientist (Agricultural Statistics) at the ICAR-Directorate of Weed Research (ICAR-DWR), Jabalpur, Madhya Pradesh, India. She earned her M.Sc. and Ph.D. in Agricultural Statistics from the ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi. Her doctoral research focused on Small Area Estimation for Spatially Correlated Data Using a Bayesian Approach. With over 16 years of experience in applied agricultural statistics, Dr. Gharde has made significant contributions to statistical modelling and its applications in agriculture. Her research spans the modelling and prediction of the future distribution of alien invasive weeds, modelling the establishment of *Zygogramma bicolorata* for the biological control of *Parthenium hysterophorus*, assessment of yield and economic losses caused by weeds in India, and the development and application of Bayesian methods and small area estimation techniques.



PROF. SHARVARI SHUKLA
Symbiosis International University

Sharvari Shukla is a Professor and Director at Symbiosis Statistical Institute, India. Sharvari is a thought leader and well regarded influencer working at the intersection of Medical Statistics and Data Science. In a prior assignment she was leading clinical research and clinical data management programme at the Savitribai Phule Pune University. She has served as a Statistical Consultant to the Pharmaceutical Industry. She has handled various responsibilities as DSMB and Ethics Committee member working on Clinical Trials. Earlier in her career she was also a Visiting Researcher at Exeter Medical School with a young researcher award. Her interest in translating research into practice led her to undergo training in epidemiology, epigenetics and clinical trials. She earned her PhD in Population Statistics and Master's Degree with specialization in Biostatistics from Pune University.



Credit Risk modelling and loss forecasting*Sumaiya Sande, Wells Fargo*

Credit risk modelling and loss forecasting are essential tools that help financial institutions assess the likelihood of borrower default and estimate potential financial losses. These models play a critical role in lending decisions, capital planning, regulatory compliance, and overall risk management. Traditionally, statistical techniques such as logistic regression and scorecard models have been widely used for credit risk assessment. However, the growing availability of financial data and advances in artificial intelligence have led to the adoption of machine learning models that offer improved predictive performance. This presentation provides an overview of the key concepts of credit risk modelling, including Probability of Default (PD), Exposure at Default (EAD), and Expected Credit Loss (ECL). It also compares traditional and modern forecasting techniques, discusses the challenges associated with model development and validation, and highlights the importance of explainable and robust models in meeting regulatory requirements.

time-dependent modeling of oncology disease progression during pandemic disruptions, we demonstrate how advanced statistical thinking refines public health governance. By deploying accelerated failure time (AFT) frameworks and auto-regressive structures, we capture non-linear transmission dynamics and systemic vulnerabilities. Further, this approach transforms raw clinical data into optimized disease models and AI-driven early-warning systems, building a resilient, self-reliant healthcare infrastructure.

The Statistical Foundations of Modern Ad Tech*Swastika Mohapatra, Google*

The rapid growth of digital advertising has led to the development of complex decision-making systems that operate at massive scale and under considerable uncertainty. From selecting which advertisement to display, to measuring campaign effectiveness and optimizing advertiser spend, statistical methods remain fundamental to the design and evolution of these platforms.

This talk will explore the role of statistics in modern ad technology through a selection of representative industry applications such as real time bidding, conversion prediction and multi touch attribution. It will highlight the practical statistical challenges encountered in real world systems and methodologies built to address them. The presentation aims to provide a practitioner's perspective on the intersection of statistics, large-scale data and online advertising.

Leveraging Data Science for species distribution modelling and risk assessment of alien invasive weeds*Yogita Gharde, ICAR-Directorate of Weed Research (ICAR-DWR), Jabalpur, India.*

NA

Statistical thinking to Policy Impact-Navigating Disease Modeling & AI for Viksit Bharat*Sharvari Shukla, Symbiosis Statistical Institute, Symbiosis International University*

The study synthesizes tertiary care insights with advanced epidemiological forecasting to drive data-driven policymaking for Viksit Bharat. Utilizing robust empirical datasets including clinical predictors of ICU requirement among COVID-19 patients and



Session Info



S212

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED PANEL

October 6th, 14:00 - 15:00 UTC

Career Perspectives from Industry in the U.S.

USE SCHED TO ACCESS THE ZOOM ROOM

This session will feature three 10-minute presentations from women working across industry sectors, including transportation, pharmaceuticals, and government research. Speakers will share insights into their career paths, current roles, and day-to-day responsibilities. The presentations will be followed by a moderated discussion exploring key topics such as career progression, mentorship, opportunities for growth, and leadership, offering attendees practical perspectives on navigating and advancing in the field of statistics and data science.

ORGANIZER: Cynthia Bland, RTI International**CHAIR:** Sophie Salem, RTI International

Speaker Bios

S212



MS. SONALI GARG
Xilio Therapeutics

Sonali Garg is an accomplished statistical programming leader with nearly 20 years of experience across the pharma, biotech, and CRO industries. She holds a graduate degree in biostatistics and an undergraduate degree in statistics. Throughout her career, Sonali has guided numerous compounds through complex global submission, approval, and post-approval processes. She has prepared multiple successful submissions to the FDA and EMA, navigated regulatory audits, and worked directly with regulators and senior management. A CDISC Subject Matter Expert, she provides critical oversight for core regulatory documents. Her extensive therapeutic expertise spans Ophthalmology, Oncology Solid Tumors, and Hematological Cancers. A strategic leader skilled in CRO oversight and vendor management, Sonali excels at aligning cross-functional teams while translating complex outputs for diverse audiences.

<https://www.linkedin.com/in/christian-douglas04/>



DR. CHRISTIAN DOUGLAS
RTI International

Dr. Christian E. Douglas is a Senior Research Statistician at RTI International with more than 14 years of experience supporting clinical trials and large-scale observational studies. Her expertise includes longitudinal data analysis, adaptive trial design, power and sample size estimation, and statistical programming in SAS and R. Since 2021, she has served as Project Director for the NIH ECHO Data Analysis Center, leading biospecimen data management, public-use dataset development, and statistical analyses for a national longitudinal cohort study of child health and development. She has led statistical efforts for federally funded clinical trials, including the ACTIV-5 BET COVID-19 trial and the RECOVER-LC Long COVID trial. Dr. Douglas earned her PhD in Biostatistics from the University of North Carolina at Chapel Hill and focuses on advancing public health through rigorous, collaborative statistical science.

<https://www.linkedin.com/in/ashley-asmus-a8119a93/>



DR. ASHLEY ASMUS
RSG Inc.

Ashley Asmus is a researcher, project manager, and data scientist with over nine years of experience working with complex datasets, including six years focused on travel survey data in both public and private sectors. Prior to joining RSG Inc., she served as lead researcher and deputy project manager for the Twin Cities Metropolitan Council's multi-year travel survey program and managed a regional origin-destination survey of more than 25,000 transit riders.

She is committed to open, reproducible science and develops transparent, well-documented analytical workflows grounded in version-controlled code and rigorous quality control. Ashley is skilled at translating complex technical concepts for broad audiences and creating documentation that enables replication. She holds a Ph.D. in Quantitative Biology and applies her background in biodiversity and global change to analyzing human travel behavior.



Session Info

S213

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 15:00 - 16:00 UTC

Causal Outside the Classroom: How is causal inference used in day-to-day working environments?

USE SCHED TO ACCESS THE ZOOM ROOM

Causal inference is a term that has become a buzzword and a popular field of research in many fields recently, and arises often in statistics. Causal methodologies can sound overly complex to many collaborators, and others may fixate on the need to have causal results from all possible studies. New researchers and scientists are left with questions of how often to employ these newer methods and frameworks of thinking, when it is appropriate to do so, and what results can actually be drawn from causal analyses.

In this session, we will hear from four professionals who use causal inference in their day-to-day practice. First, Dr. Candice Johnson will discuss whether certain causal studies are inherently gendered by examining hazards in the workforce in male-dominated fields. Dr. Lucy McGowan will talk about whether interaction effects are feasible, clinically significant, and better than subgroup-specific effects in real-world trials. Dr. Maria DeYoreo will explore inferential approaches to estimate causal effects from a study on the effect of state-funded family planning policies on birth outcomes. Finally, Dr. Lu Wang will discuss applying a causal framework to personalized healthcare, demonstrating how causal inference can support flexible, patient-centered decision-making in routine clinical practice.

ORGANIZER: Jamie Joseph, Henry Ford Health**CHAIR:** Jamie Joseph, Henry Ford Health**SPONSOR:** STATA

Speaker Bios

S213



DR. CANDICE JOHNSON
Michigan State University

<https://epibio.msu.edu/faculty/johnson.html>

Candice Johnson is Assistant Professor in the Department of Epidemiology and Biostatistics at Michigan State University. She studies how work-related exposures and policies affect the health of workers and their families, with a focus on women and pregnant women in the workforce. Dr. Johnson also conducts research on causal inference topics related to misclassification, the measurement and use of race in epidemiology and clinical guidelines, and studying women in the workforce. The goal of her methodologic research is to provide epidemiologists with easy-to-use tools and guidance that they can apply to their research.



DR. LUCY D'AGOSTINO MCGOWAN
Wake Forest University

<https://www.lucymcgowan.com/>

Lucy D'Agostino McGowan is an associate professor in the Department of Statistical Sciences at Wake Forest University. She received her PhD in Biostatistics from Vanderbilt University and completed her postdoctoral training at Johns Hopkins University Bloomberg School of Public Health. Her research focuses on causal inference, statistical communication, analytic design theory, and data science pedagogy. Dr. D'Agostino McGowan was the 2023 chair of the American Statistical Association's Section on Statistical Graphics and can be found blogging at livefreeordichotomize.com, on Twitter @LucyStats, and podcasting on the American Journal of Epidemiology partner podcast, Casual Inference.



DR. MARIA DEYOREO
RAND

https://www.rand.org/about/people/d/deyoreo_maria.html

Maria DeYoreo is a senior statistician at RAND. Her substantive interests include health care quality and performance measurement, Medicare Advantage, hospice care, and maternal/child health. Her methodological interests include causal inference and quasi-experimental methods, Bayesian modeling, multi-level modeling, missing data imputation, and data fusion. She co-leads a team that works with the Centers for Medicare & Medicaid Services to generate the Medicare Advantage Star Ratings (quality performance ratings for health plans). DeYoreo received her B.S. in mathematical sciences from the University of California, Santa Barbara, and her Ph.D. in statistics from the University of California, Santa Cruz, where she developed flexible Bayesian nonparametric models for ordinal regression. She completed postdoctoral research at Duke University, focused on missing data imputation and data fusion.



DR. LU WANG
University of Michigan

<https://sph.umich.edu/faculty-profiles/wang-lu.html>

Dr. Lu Wang received her PhD in Biostatistics from Harvard University in 2008 and joined the faculty at the University of Michigan in the same year. Her research focuses on statistical methods for evaluating dynamic treatment regimes, personalized health care, nonparametric and semiparametric regressions, missing data analysis, functional data analysis, and longitudinal (correlated/clustered) data analysis. Dr. Wang is an ASA fellow and currently serves as an Associate Editor of Biometrics. She also serves as the Core Co-Leader of the Integrated Health Sciences Core (IHSC), Michigan Center on Lifestyle Environmental Exposures and Disease (M-LEEAD).



Is causal inference gendered? Examples from a male-dominated field*Candice Johnson, Michigan State University*

In occupational epidemiology, we study how workplace exposures affect health. Historically, our field has focused on hazards found in male-dominated jobs, such as coal miners or construction workers, while neglecting hazards in female-dominated jobs. As a result, the methods and concepts used in our field in pursuit of causal inference were developed almost exclusively based on studies of men. In the United States, women and people of other gender identities now participate in the workforce at high rates and are of increasing interest to occupational epidemiologists. Can we use our field's traditional methods and concepts to study these populations, or are our methods and concepts gender-specific? In this presentation, I will present examples from our research in which the methods and concepts used routinely in our field did not work as expected when applied to studies of women. These results help us to better understand how gender can play an underappreciated role in causal inference.

Patients Don't Care About the Interaction Term (Should We?)*Lucy D'Agostino McGowan, Wake Forest University*

Powering a trial to detect subgroup-by-treatment interactions is, in most realistic scenarios, simply not feasible. At the design stage, investigators presume clinical equipoise, that is they do not expect treatment to harm any group of patients. The overall effect is the weighted average of subgroup-specific effects and the expectation of clinical equipoise sharply constrains how different those effects can be. Under these assumptions, the sample sizes required to detect interactions are dramatically larger than those needed to detect the overall treatment effect. In a trial designed for 80% power on the overall effect, detecting an interaction of comparable magnitude would require at least four times the sample size, and even under ideal conditions, interaction test power hovers around 29%. However, power to detect a subgroup-specific effect is not the same as power to detect an interaction, and these reflect fundamentally different scientific goals. Typically, a patient does not care whether their treatment effect is statistically distinguishable from another subgroup's, they care whether there is a clinically and statistically meaningful benefit within their own subgroup. Meaningful, valid estimates of subgroup-specific effects are achievable without relying on formal interaction tests, and pre-specifying subgroups and reporting within-subgroup estimates directly is both statistically sound and more aligned with how patients and clinicians actually use trial results.

The effect of state-funded family planning policies on birth outcomes*Maria DeYoreo, RAND*

State-funded family planning (SFP) programs extend reproductive health coverage — including contraception, STI screening, and preventive care — to low-income women excluded from Medicaid due to income, immigration status, or other factors.

Despite their reach, evidence on the causal impact of these programs on reproductive and economic outcomes remains limited. We estimate the effect of SFP adoption on birth rates, contraceptive access, and state and federal health care spending across five states (IL, OR, SC, TX, UT) using a stacked difference-in-differences design with explicit matching on competing policy environments. We explore different inferential approaches to account for few clusters (states), and assess sensitivity to modeling assumptions. This study provides the first comprehensive multi-state assessment of SFP programs, enabling comparison of alternative program design features and informing future state investment decisions.

Causal Inference in Clinical Practice: Learning Personalized Treatment Decisions from Data*Lu Wang, University of Michigan*

Personalized healthcare often requires balancing competing objectives, such as improving survival while minimizing treatment burden. Causal inference provides a principled framework for learning individualized treatment strategies from data, but in practice there is often more than one clinically acceptable option. In this talk, I introduce the concept of a tolerant dynamic treatment regime, which provides a set of feasible decision rules that achieve outcomes within a prespecified tolerance of the optimum. I will present an interpretable tree-based causal learning approach for estimating the optimal decision rules depending on the decision-maker's preferences and illustrate its application to sequential chemotherapy decisions for advanced prostate cancer. This work demonstrates how causal inference can support flexible, patient-centered decision-making in routine clinical practice.



Session Info

S214

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED PANEL

October 6th, 17:00 - 18:00 UTC

Beyond Automation: Using Generative AI to Elevate Critical Thinking, Personalization, Engagement, and Analytics Instruction

USE SCHED TO ACCESS THE ZOOM ROOM

Generative AI is rapidly reshaping the landscape of statistics and business analytics education, offering new opportunities to improve course design, assignment development, and feedback processes. This panel will share practical strategies for integrating AI into teaching workflows while maintaining statistical rigor and fostering student understanding.

We highlight approaches for AI-assisted case and assignment design, including adaptive scenarios in which students pose questions, request instructor-controlled synthetic data, and iteratively refine their analyses. These methods promote deeper engagement with modeling choices, assumptions, and critical thinking-core goals of statistics education.

The panel also demonstrates how AI can support assessment and feedback, from automated classwide test analysis to personalized guidance that helps students interpret results and correct misconceptions.

Attendees will leave with concrete examples, implementation tips, and considerations for responsible use of AI tools in statistics and business analytics classrooms.

ORGANIZER: Jessica Kohlschmidt, Caucus for Women in Statistics and Data Science

CHAIR: Jessica Kohlschmidt, Caucus for Women in Statistics and Data Science



Speaker Bios

S214



DR. JOHN DRAPER
The Ohio State University

John Draper is an Associate Professor of Clinical Operations and Business Analytics in the Fisher College of Business at The Ohio State University. He has a PhD and MS in statistics from Ohio State as well as BS degrees in mathematics and statistics from Florida State University. His extensive experience in education at Ohio State includes courses statistical theory, business analytics, engineering, operations, six sigma, statistical computing/data science, and sports analytics at both the graduate and undergraduate level as well as biostatistics courses for graduate-level students in the College of Public Health and Dentistry. He currently serves as the Co-academic Director of the Specialized Masters of Business Analytics and the Academic Director of the Executive Education Master Black Belt.



DR. JESSICA KOHLSCHMIDT
Caucus for Women in Statistics and Data Science

Dr. Jessica K. Kohlschmidt is a distinguished biostatistician, educator, and leader in the statistical community. A first-generation college student, her early passion for math and teaching evolved into a dedicated career in statistics after discovering its real-world applications. She has built a multifaceted career that combines teaching, high-impact clinical research, and significant leadership roles in professional organizations. For 12 years, she applied her expertise to leukemia outcomes research and she has been a college educator since 1999. Dr. Kohlschmidt is deeply committed to mentorship and outreach, particularly encouraging young women to pursue statistics, and holds key leadership positions in several statistical organizations, including the American Statistical Association (ASA).



MR. GEORGE RECK
Babson College

Mr. Recck has taught at Babson College since 1984. Prior to that, Mr. Recck worked his way toward his undergraduate and MBA degrees from Babson while working in Babson's Computer Center. After receiving his MBA from Babson in 1984, Mr. Recck was promoted to Director of Academic Computer Services and began his teaching career shortly thereafter. Mr. Recck founded Total Information, Inc., a consulting firm in 1991 that continues to this day providing a variety of clients a host of information services. Mr. Recck became the Director of the Math Resource Center at Babson in 2005 and is also an Associate Professor of Practice in the Math, Analytics, Science, & Technology division. Mr. Recck has taught a variety of Business Analytics courses at all levels. Mr. Recck has also published in the areas of teaching business analytics and using analytics in sports.



MS. CHERYL EDELMANN
University of Dayton

Cheryl Edelmann is a professor of Business Analytics in the Department of MIS, OSC, and Business Analytics at the University of Dayton. Cheryl educates future business professionals in the area of statistics, helping them develop the skills needed to succeed in today's data-driven world. Her enthusiasm for teaching and passion for guiding students as they discover their vocations are central to her professional work.



Session Info

S215

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 18:00 - 19:00 UTC

From Curiosity to Confidence: Young Women's Journeys in Statistics and Data Science

USE SCHED TO ACCESS THE ZOOM ROOM

Organized by Florence Nightingale Day (FNDay), a 501(c)(3) nonprofit organization.

The Florence Nightingale Day for Statistics and Data Science (FNDay) was established to inspire and support students at an early age as they explore statistics, data science, and related quantitative disciplines through mentorship, educational opportunities, and community engagement. This session showcases the impact of FNDay through the experiences of four student presenters, including two high school students and two undergraduate students. Three of the speakers are past FNDay participants, while the fourth will participate in an upcoming FNDay event.

The students will present projects that demonstrate the power of statistics and data science to address important challenges in health and education. Their research explores questions such as how data can help us reveal patterns of cellular aging, better understand colon cancer risk, identify factors associated with the progression of colorectal polyps to cancer, and examine whether investments in school districts are linked to improved student achievement. Together, these presentations highlight how early exposure to quantitative thinking, meaningful mentorship, and supportive learning environments can foster curiosity, build confidence, and encourage young women to pursue research and academic pathways in statistics, data science, and related fields.

ORGANIZER: Xiaokui Mo, The Ohio State University; FNDay 501(c)(3)

CHAIR: Mia Yu, The Ohio State University



Speaker Bios

S215



CASSIE YUAN
The Ohio State University

Cassie Yuan is a third-year Data Analytics major at The Ohio State University, specializing in Public Health and Biomedical Informatics. She is actively involved in the Mount Leadership Society Scholars program and the Big Data Analytics Association. In summer 2025, she completed a research internship with Ohio State's Department of Biomedical Informatics, where she investigated racial and ethnic disparities in cervical cancer screening and patient use of online medical records. In summer 2026, she worked with the Ohio Department of Education and Workforce, applying R-based data analysis to evaluate school district expenditure patterns across the state. Cassie is passionate about using data science to understand complex systems, support evidence-based decision-making, and improve public health outcomes. A longtime supporter of FNday, she has participated in the program since eighth grade and has continued her involvement as an undergraduate student representative. She previously served on the FNday student panel and has actively contributed to outreach efforts, helping inspire students to explore opportunities in data science, informatics, and biomedical research.



BIANCA EHLING
High School Student

Bianca is a senior in High School with a strong interest in bioinformatics, computational biology, and the use of artificial intelligence to advance cancer research. She conducts research at The Ohio State University, where she applies deep learning methods to investigate senescence programs in retinoblastoma, and recently participated in Ohio State's CREATES High School Program, an 8-week funded cancer research experience. Bianca presented her research at the 2026 International Conference on Intelligent Biology and Medicine and received a Student and Trainee Travel Award. She also participated in FNday at The Ohio State University in 2025. In addition to her research, Bianca is passionate about science education and outreach. She created Gene Regulation, a Science Olympiad event focused on genetics, epigenetics, and data science that has been implemented in 23 states and reached more than 600 students. She also founded a coding program for elementary school students to introduce fundamental programming concepts.



CLAIRE ERDELYI
The Ohio State University

Claire Erdelyi is a senior in The Ohio State University College of Arts and Sciences from Marietta, Ohio, majoring in Data Analytics. She currently works as a Data Analyst in Ohio State's Office of Technology and Digital Innovation (OTDI), where she supports data-driven decision-making and analytics initiatives. Claire is also active in the International Affairs program and served on the 2026 DataFest Student Board. During the summer of 2026, Claire completed a research internship in the Department of Biomedical Informatics, where she gained hands-on experience applying statistical and data analytic methods to biomedical research. Her work focused on understanding patterns of colorectal cancer development through quantitative data analysis. Following graduation, she plans to pursue a Master's degree in Data Science. Claire has been invited to present her colorectal cancer research as a poster at FNday 2026 and will also serve as a member of the student panel. She is passionate about leveraging data science and analytics to address complex challenges in healthcare and biomedical research.



ANNABETH PAN
High School Student

Annabeth Pan is a senior at Wellington School in Columbus, Ohio. She is the co-founder and captain of her school's Science Olympiad team, leading the team to numerous regional medals and state tournament qualification. At the National Robotics Challenge, Annabeth and her teammates received the Honda Innovation Award for designing an autonomous navigation vehicle. This summer, she participated in the Summer Science Program, conducting research in astrophysics, and also interned with the BMBL Bioinformatics Lab at The Ohio State University, where she gained experience in computational and biomedical research. Annabeth presented her research at FNday 2025, earning the Second Best Poster Award for her outstanding work. Passionate about science, engineering, and data-driven discovery, Annabeth enjoys exploring interdisciplinary applications of technology to solve complex problems and hopes to pursue further study and research at the intersection of computational and biological sciences.



Prediction of Senescent Programs in Retinoblastoma*Bianca Ehling, High School Student*

Retinoblastoma is the most common pediatric intraocular eye cancer, affecting approximately 9,000 children globally each year. The objective of this research was to computationally characterize senescent cells (SnCs) in retinoma (RE) and retinoblastoma (RB). We also wanted to investigate any differences between senescence programs in RE and RB, and also if cellular senescence is associated with RE-RB development. It has been shown that SnCs are enriched in RE. Also, CDKN1A (p21), CDKN2A (p16), RBL2 (p130) are associated with senescence in RE/RB. Yet, the specific senescence programs in retinoma and how they could affect the transition from retinoma to retinoblastoma need to be further explored. Our group has developed a new deep learning tool to investigate cell type-specific senescent cells and genes, named DeepSAS. It is the first tool that can deliver de novo cell type senescent programs and has overcome existing tools in identifying senescent cells especially for human patient samples. Using DeepSAS, we performed analysis on public single-cell RNA-seq data from intraocular and extraocular retinoblastoma patient samples. Findings include that 10% of SnCs were predicted to be found in RE and RB, which was unexpected. This could be attributed to heterogeneity in RB. Senescence key markers (CDKN1A, CDKN2A, and RBL2) showed higher expression levels in SnCs than non-SnCs, and lower in RB than RE. Senescence associated genes (SnGs) predicted in RE were mostly related to DNA-damage-associated senescence and growth arrest. In RB, SnGs serve a function more characteristic of malignant, proliferative, and aggressive cancers (tumor-supportive program). Expression of SnGs is highly correlated with CDKN1A and CDKN2A in RE and RB, while weakly correlated with RBL2.

Lifecycle of Methylation in Colorectal Cancer*Claire Erdelyi, The Ohio State University*

Previous studies have demonstrated that aberrant DNA methylation accumulates in colorectal cancer development, suggesting that progressive epigenetic alterations may distinguish normal tissue, adenomas, and colorectal carcinomas. Therefore, I hypothesized that DNA methylation patterns vary significantly across these stages of colorectal cancer progression. To investigate this hypothesis, I employed two analytical approaches: (1) the calculation of a global methylation for each sample by summing methylation values across CpG sites, followed by a one-way analysis of variance (ANOVA) to test for overall differences among disease stages and post hoc pairwise comparisons to identify which specific stages differed significantly; and (2) the generation of volcano plots to demonstrate differential methylation at individual CpG sites between disease states.

The analysis revealed significant methylation differences between adjacent stages of colorectal cancer progression and between samples from distinct disease states stratified by attributes such as sex and cancer stage. In contrast, comparisons among samples within the same disease state identified few significant methylation differences. Furthermore, differential methylation was observed at the majority of CpG sites when comparing samples across disease states.

Clustering Ohio Public School Districts on Expenditures*Cassie Yuan, The Ohio State University*

Ohio's public school districts vary substantially in size, geographic location, student demographics, and educational needs, creating complex challenges for equitable and effective resource allocation. Understanding how districts distribute financial resources is essential for evaluating spending practices and informing education policy. This project examines patterns in per-pupil expenditures across Ohio school districts by analyzing both classroom-related expenditures, such as instruction and student support services, and non-classroom expenditures, including administration, operations, and transportation.

Using statistical methods and K-medoids clustering, districts were grouped according to similarities in their spending profiles. The analysis identified several distinct expenditure patterns, highlighting meaningful variation in how districts prioritize and allocate resources. Associations between spending patterns and district characteristics, including enrollment size, socioeconomic factors, and geographic context, were also explored. While some relationships were observed, spending patterns showed only modest correspondence with district attributes and limited alignment with existing state district typology classifications.

These findings demonstrate the value of data science techniques for uncovering hidden structure within complex educational finance data. By providing a data-driven perspective on resource allocation, this work offers insights that may support policymakers, administrators, and researchers in interpreting spending practices and evaluating the effectiveness of current classification systems across Ohio's public school districts.

High-Throughput RNA Sequencing Data Analysis in Colorectal Cancer*Annabeth Pan, High School Student*

Cancer Colorectal cancer begins in the colon or rectum. This disease involves major genetic alterations in the affected cells, which can be detected with differential gene analysis. This technique compares RNA counts—how often each gene is transcribed—between healthy and cancerous tissues to identify differentially expressed genes (DEGs), which can reveal pathways for developing targeted therapies. We performed Variance stabilizing transformations, Principal Component Analysis clustering, and the DESeq2 package to analyze previously collected raw bulk RNA-seq data. Between the cancerous and non-cancerous samples, it was found that Rps12-ps3, Ccl21b, and Ccl21d were consistently most overexpressed in colorectal cancer, and Gsdmc2, Ces1g, and Hao2 the most underexpressed. Further exploration can determine causality or create targeted drug therapies based on these identified genes.



Session Info

S216

TECHNICAL STATISTICAL SESSION

INVITED PANEL

October 6th, 20:00 - 21:00 UTC

Frontiers in Statistical Modeling: High-Impact Doctoral Research from Argentina (Fronteras en la Modelización Estadística: Investigaciones Doctorales de Alto Impacto desde Argentina)

USE SCHED TO ACCESS THE ZOOM ROOM

Statistical training in Argentina is characterized by quantitative rigor and methodological innovation. Postgraduate programs of excellence, like the PhD in Statistics at the Universidad Nacional de Rosario, build this foundation. This session showcases this training through four doctoral theses that propose cutting-edge solutions for high dimensionality, spatial dependence, incomplete data, and outliers.

Attendees will learn about:

Associative Mapping: Verónica Lac Prugent combines Multivariate Elastic Net models to address quantitative and qualitative responses in genetic resources, ensuring the empirical robustness of associations.

Multi-Environment Trials: Julia Angelini addresses the limitations of the Sites Regression (SREG) model by proposing an Expectation-Maximization (EM) based imputation method and robust variants to reduce the impact of outliers.

Spatial Prediction: Mariel Lovatto introduces a semiparametric spatial autoregressive model that captures spatial dependence without relying on strict parametric covariance, resulting in improved predictive performance.

Zero-Inflated Data: María José Llop introduces a robust EM-type estimation approach for the partially linear zero-inflated Poisson (PLZIP) regression model, effectively handling extreme observations in complex relationships.

This panel highlights the level of Argentine science and its capacity to generate statistical knowledge with global impact.

ORGANIZER: María Inés Stimolo, Instituto de Estadística y Demografía, Universidad Nacional de Córdoba

CHAIR: María Inés Stimolo, Instituto de Estadística y Demografía, Universidad Nacional de Córdoba



Speaker Bios

S216



DR. VERÓNICA PATRICIA LAC PRUGENT
Universidad Nacional del Comahue

<https://www.researchgate.net/profile/Veronica-Lac-Prugent>

She holds a Bachelor's degree in Statistics, a Specialization in Epidemiology, and is a PhD in Statistics. She works as a Professor and researcher in the Department of Statistics at the Faculty of Economics and Administration of the Universidad Nacional del Comahue, where she also serves as Deputy Director of the Department of Statistics and Director of the Observatory of Educational and Salary Statistics. Her work focuses on statistical modeling, multivariate analysis, and statistical learning, with applications in health, education, and agricultural sciences. She has directed and participated in research projects, and has authored scientific publications and presented at both national and international conferences.



DR. JULIA ANGELINI
Universidad Nacional de Rosario

<https://www.researchgate.net/profile/Julia-Angellini>

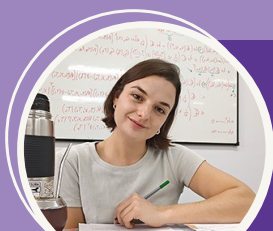
She holds a Bachelor's degree in Statistics, a Specialization in Bioinformatics, and a PhD in Statistics. She serves as a Professor at the Faculty of Economic Sciences and Statistics of the Universidad Nacional de Rosario and as a senior advanced data scientist at NaranjaX. Currently, her work focuses on statistical modeling and multivariate analysis from a dual perspective: professional practice in the industry and university teaching. Her career includes participation in research projects, as well as a track record of scientific dissemination that encompasses articles in peer-reviewed journals and presentations at national and international academic conferences.



DR. MARIEL LOVATTO
Universidad Nacional del Litoral (UNL) - CONICET

<https://bicyt.conicet.gov.ar/fichas/p/mariel-guadalupe-lovatto>

She holds a degree in Mathematics from the Universidad Nacional del Litoral, and a Master's and a PhD in Statistics from the Universidad Nacional de Rosario. She currently holds a postdoctoral fellowship from CONICET and serves as Head Teaching Assistant at the Faculty of Chemical Engineering of the Universidad Nacional del Litoral. Her research focuses on the development of nonparametric and semiparametric models for spatial data analysis. She has presented her research findings at national conferences and has authored publications in national and international scientific journals.



DR. MARÍA JOSÉ LLOP
Universidad Nacional del Litoral

She holds a PhD in Statistics from the Universidad Nacional de Rosario and a Bachelor's degree in Applied Mathematics from the Universidad Nacional del Litoral. She is currently a postdoctoral fellow at CONICET; she serves as a Professor at the Faculty of Chemical Engineering and as Head Teaching Assistant at the Faculty of Engineering and Water Sciences of the UNL. Her research focuses on semiparametric statistics, robust methods, regression models for zero-inflated data, and statistical learning. She has presented her findings at national and international conferences and has published her work in international scientific journals.



Mixture-based Nonparametric Estimation of Spatial Covariance Functions with Applications to HIV Key Population Size Estimation across Sub-Saharan Africa

Manushi Siriwardana, Penn State

Consistent data on the sizes of key populations, such as female sex workers (FSWs), are often scarce, particularly at the sub-national level. Accurate size estimates are critical to effectively allocate resources and achieve HIV targets. Since FSW population sizes may be spatially correlated across areas, models that account for spatial dependence can improve estimation. An important component of such models is the covariance function, which characterizes the spatial dependence structure of the underlying process. In this work, we study spatial covariance functions to estimate FSW population sizes in Sub-Saharan Africa (SSA). Parametric covariance estimation can suffer from model mis-specification, potentially leading to inefficient or biased predictions. We therefore develop a robust non-parametric approach for estimating the covariance function of a stationary isotropic process in R^d . We focus on a class of covariance functions that are valid in all dimensions, which includes popular kernels such as the exponential and Mat'ern kernels. Leveraging the fact that such covariance functions can be represented as infinite mixtures of scaled Gaussian kernels, we propose two estimation methods: weighted least squares and nonparametric maximum likelihood estimation to estimate the mixing measure of scaled Gaussian kernels. We also develop computationally efficient methods to solve these optimization problems using non-negative least squares and second-order descent updates.

Influenza hospitalization tracking using internet search data and ILI surveillance

Hanqi Shi, Swarthmore College

For influenza control and public health decision-making, timely estimates of influenza hospitalization rates are essential. Unlike traditional influenza surveillance, which is reported with a delay of one to three weeks, Internet search data provide timely signals of ongoing influenza activity. Here we present a statistical framework for real-time tracking of influenza hospitalization rates in the United States. Our approach builds on the ARGO and ARGOX frameworks by combining Google Trends data with lagged influenza-like illness (ILI) surveillance through dynamically trained regularized regression models. Using hospitalization data from the Centers for Disease Control and Prevention's National Healthcare Safety Network (NHSN), our method outperforms models that rely on Internet search data alone. The proposed framework demonstrates that integrating multiple surveillance signals can improve real-time influenza hospitalization tracking and provides a foundation for future joint modeling of ILI activity and hospitalization rates.

The Impact of Recent State Staffing Minimum Mandates on Nursing Home Residents

Angira Mondal, Perelman School of Medicine, University of Pennsylvania

To ensure adequate staffing, some U.S. states have instituted

minimum staffing mandates for nursing homes, defined by the Hours Per Resident Day (HPRD). This study explores the impact of these state-level staffing mandates on staffing HPRD and nursing home resident outcomes such as proportion of residents with stage 2 or higher-pressure ulcers, proportion of residents receiving inappropriate antipsychotic medication and the mortality rate. Heterogeneity analysis was completed by dementia status. Using CMS administrative data, we conduct employ staggered difference-in-difference methodology to study the effect of new or updated mandates implemented by 7 states between 2018 to 2023 (CA, AR, MA, CT, NY, RI and PA). We observe that the staffing minimum leads to an increase in the total staffing HPRD, but the average masks considerable state variation, ranging from no effect in AR to significant increased effect in CA. The resident outcomes also generally improve, especially in states where we observe greater staffing increase post the implementation of the regulations. Our findings show that staffing adequacy contributes to poor resident outcomes in US nursing homes.

From Raw Wearable Data to Scientific Inference: A Reproducible Pipeline for Processing Commercial Step Count Data in Aging Cohorts

Nannan Bo, Department of Biostatistics and Bioinformatics, Duke University School of Medicine; Duke Aging Center, Duke University School of Medicine; incoming PhD student, Quantitative Biomedical Sciences Program, Dartmouth College

Commercial wearable devices offer an unprecedented opportunity to objectively measure free-living physical activity at scale. However, raw data generated by consumer devices are rarely analysis-ready. These data are often recorded in epochs (fixed time intervals) and may contain irregular timestamps, missing intervals, and limited wear-time information, making processing decisions critical to downstream scientific inference.

We have developed a transparent, reproducible standard operating procedure (SOP) to transform Garmin epoch-level step-count data into participant-level analytic variables in the STRRIDE-PD Reunion Study, a long-term follow-up of older adults originally enrolled in a supervised exercise intervention trial. The pipeline standardizes timestamps, reconstructs epoch grids, infers wear time, identifies valid observation days, and generates participant-level activity summaries with quality control and sensitivity analyses.

Among 67 participants, the pipeline produced robust estimates of daily step-count metrics across alternative processing criteria. Models using wearable-derived step counts identified associations with cardiometabolic and fitness outcomes that were not detected using self-reported physical activity, demonstrating how measurement approach can influence scientific conclusions.

This work highlights that processing commercial wearable data is not merely data cleaning but a measurement problem requiring transparent, reproducible analytic decisions.



Session Info

S217

TECHNICAL STATISTICAL SESSION

CONTRIBUTED SESSION

October 6th, 21:00 - 22:00 UTC

Statistical Methods and Applications 2

USE SCHED TO ACCESS THE ZOOM ROOM

Mixture-based Nonparametric Estimation of Spatial Covariance Functions with Applications to HIV Key Population Size Estimation across Sub-Saharan Africa | Influenza hospitalization tracking using internet search data and ILI surveillance | The Impact of Recent State Staffing Minimum Mandates on Nursing Home Residents | From Raw Wearable Data to Scientific Inference: A Reproducible Pipeline for Processing Commercial Step Count Data in Aging Cohorts

ORGANIZER: IDWSDS Program Committee**CHAIR:** Adeeba Tak, Stanford University

Speaker Bios

S217



MISS MANUSHI SIRIWARDANA

Penn State

Manushi Siriwardana is a Ph.D. student in Statistics at Penn State. She will be entering her fourth year of the Ph.D. program in Fall 2024. Her research focuses on spatial statistics, public health, and computational statistics. She is interested in developing statistical methodology and computational tools for analyzing complex data, with applications to public health and related fields.



MS. HANQI SHI

Swarthmore College

Hanqi Shi is an undergraduate student at Swarthmore College majoring in Statistics and Computer Science. Her research interests include statistical modeling, machine learning, and interdisciplinary applications.



ANGIRA MONDAL

University of Pennsylvania

Angira Mondal is a Senior Statistical Analyst at the Perelman School of Medicine, University of Pennsylvania, where she works in the Policy and Economics of Disability, Aging, and Long Term Care (PEDAL) Lab. Her research focuses on complex healthcare data, including claims analysis, leveraging wearable device metrics to enhance health outcomes, and evaluating Centers for Medicare & Medicaid Services (CMS) data to assess nursing home staffing policies. Angira holds a master's degree in statistics and a bachelor's degree in economics and statistics from the University of Minnesota.



MISS NANNAN BO

Duke University; Dartmouth College

Nannan Bo recently completed her Master of Biostatistics at Duke University and will begin her PhD training in the Quantitative Biomedical Sciences program at Dartmouth College in Fall 2026. Her research interests span clinical trial design, Bayesian methods, measurement error, wearable and real-world data, reproducible computational workflows, and applied statistical methods for biomedical research. She is particularly interested in developing rigorous and interpretable methods for complex health data, with growing interests in bioinformatics and multi-modal biomedical data analysis. Her current work focuses on transparent analytic pipelines for commercial wearable step-count data in aging cohorts.



My experiences as a collaborative biostatistician working on infectious disease clinical trials

Madeline Abbott, the Center for Biostatistics in AIDS Research (CBAR) at Harvard University

I am a Research Associate in the Center for Biostatistics in AIDS Research (CBAR) in the Harvard School of Public Health. Prior to joining CBAR, I completed my PhD in Biostatistics at the University of Michigan in 2024. As a graduate student, I worked on research projects with applications ranging from cancer to mobile health. Through these experiences, I realized that I enjoyed working closely with clinical collaborators to answer current scientific questions. I also realized that I was interested not only in the statistics, but also in understanding the scientific context behind the statistical work. This led me to my current position at CBAR. At CBAR, I work on clinical trials for the treatment and prevention of HIV and tuberculosis, collaborating closely with clinicians as part of the International Maternal Pediatric Adolescent AIDS Clinical Trials (IMPAACT) Network. My responsibilities range from developing safety monitoring and statistical analysis plans to reviewing case report forms and enrollment materials, as well as pursuing statistical methods research motivated by the clinical trial data. In this talk, I will describe my path to this position, my current responsibilities, and my experiences as an early-career collaborative biostatistician.

Training in a Data Coordinating Center: Fostering the Development of Highly Collaborative and Effective Early Career Biostatisticians

Megan McCabe, University of Alabama at Birmingham

Collaborative clinical research environments, and data coordinating centers (DCCs) in particular, require a diverse, interdisciplinary skillset, which is generally not the focus of graduate-level statistics or biostatistics coursework. Instead, these skills are typically learned and developed through research assistantships, internships, or other types of experiential learning opportunities.

Dr. McCabe trained in an academic DCC, the CTSDMC, and has pursued a career in academic DCCs post-graduation as the Assistant Director for Clinical Trials Development for the DATA CRU (UAB). Experiences as a graduate research assistant in a DCC were foundational for understanding the dynamic research environment associated with complex clinical research studies. At the CTSDMC, the speaker was the inaugural Network for Excellence in Neuroscience Clinical Trials (NeuroNEXT) Biostatistics Fellow and had the opportunity to gain experience working on clinical research studies from pre-award grant submission through dissemination and publication of results. In addition to academic DCC experience, she also completed two fellowships associated with FDA, as well as a co-op at a pharmaceutical company, Boehringer Ingelheim. This presentation will highlight the speaker's firsthand experience with training in a DCC environment and other experiential learning opportunities during her doctoral studies and how these experiences have facilitated her transition to faculty post-graduation.

Growing as a Collaborative Biostatistician in a Data Coordinating Center

Margaret Banker, Northwestern University

As an early-career faculty biostatistician in academic clinical research, I work in a Data Coordinating Centers (DCCs) where I collaborate with investigators to design, conduct, and analyze multicenter observational studies. This statistical collaboration involves developing statistical analysis plans, monitoring data quality, managing study data, and performing analyses that support study objectives and publications. In this talk, I will discuss experiences from two national research networks: the Glycemic Observation and Metabolic Outcomes in Mothers and Offspring (GOMOMS) study, coordinated through Northwestern University, and the Cure Glomerulonephropathy (CureGN) study, coordinated through the University of Michigan. I will highlight the diverse roles of biostatisticians in Data Coordinating Centers (DCCs) and offer an early-career perspective on the opportunities and challenges of collaborative biomedical research.

My Journey Through Clinical Trials: Experiences as a Faculty Member Working with a Data Coordinating Center

Emily Roberts, University of Iowa

As an Assistant Professor of Biostatistics at the University of Iowa, a significant proportion of my role is within our Clinical Trials Statistical and Data Management Center (CTSDMC) Data Coordinating Center (DCC). In particular, I work within the NeuroNEXT network to help design and analyze neurology-related clinical trials. I will share my experiences to date of transitioning from a graduate student to faculty member including my responsibilities within NeuroNEXT where I currently lead our pre-award team of biostatisticians.



Session Info



S218

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED SESSION

October 6th, 22:00 - 23:00 UTC

Experiences of Early Career Biostatisticians in Clinical Trial and Data Coordinating Centers

USE SCHED TO ACCESS THE ZOOM ROOM

This session will feature four early career researchers working in academic data coordinating and/or clinical trial centers. The speakers will talk about their (collaborative) biostatistician roles and responsibilities in their respective institutions, sharing about their training journeys and lessons learned along the way.

ORGANIZER: Emily Roberts, University of Iowa**CHAIR:** Emily Roberts, University of Iowa

Speaker Bios

S218



DR. MADELINE ABBOTT
Harvard University

I am a Research Associate in the Center for Biostatistics in AIDS Research (CBAR) at Harvard University. I work on clinical trials for the treatment and prevention of HIV and tuberculosis, collaborating closely with clinicians in the International Maternal Pediatric Adolescent AIDS Clinical Trials (IMPAACT) Network. Prior to joining CBAR, I completed my PhD in Biostatistics at the University of Michigan in August 2024. My dissertation research focused on the development of methods for jointly modeling intensive longitudinal and survival data from mobile health studies.



DR. MEGAN MCCABE
University of Alabama at Birmingham

Dr. Megan McCabe is an Assistant Professor in the Department of Biostatistics at the University of Alabama at Birmingham (UAB), and the Assistant Director for Clinical Trials Development for the DATA coordinating and Collaborative Research Unit (DATA CRU). She graduated with her PhD in biostatistics from the University of Iowa where she worked as a Graduate Research Assistant in the Clinical Trials and Statistical Data Management Center (CTSDMC). During her doctoral studies she completed internships with FDA CDER Office of Biostatistics and Boehringer Ingelheim. Prior to that, she obtained her MS in biostatistics at Northwestern University and worked as a statistical analyst in the Department of Preventive Medicine there. Her research focuses on the design, conduct, and analysis of clinical trials and observational studies, as well as on infrastructure and practices for data coordinating centers.



DR. MARGARET BANKER
Northwestern University

Dr. Margaret Banker is an Assistant Professor of Biostatistics and Informatics in the Department of Preventive Medicine at Northwestern University Feinberg School of Medicine. As a member of the Northwestern University Data Analysis & Coordinating Center (NUACC), she collaborates on the design, coordination, and analysis of multicenter clinical research studies. Her research focuses on the development and application of statistical methods for analyzing complex health data, with particular expertise in longitudinal data analysis, wearable devices, and functional data analysis.



DR. EMILY ROBERTS
University of Iowa

Dr. Emily Roberts is an assistant professor of biostatistics at the University of Iowa. She completed her PhD in Biostatistics at the University of Michigan in 2022, where her dissertation focused on Causal Inference Methods and Intermediate Endpoints in Randomized Clinical Trials. She enjoys developing statistical methodology that can be meaningfully applied in clinical and public health settings. Since joining the faculty at Iowa, she has been building and leading interdisciplinary research partnerships through the University of Iowa's Causal Inference Collaboratory and is involved in multicenter design and analysis of clinical trials through the Clinical Trials Statistical and Data Management Center via the Data Coordinating Center and NeuroNEXT.



My experiences as a collaborative biostatistician working on infectious disease clinical trials

Madeline Abbott, the Center for Biostatistics in AIDS Research (CBAR) at Harvard University

I am a Research Associate in the Center for Biostatistics in AIDS Research (CBAR) in the Harvard School of Public Health. Prior to joining CBAR, I completed my PhD in Biostatistics at the University of Michigan in 2024. As a graduate student, I worked on research projects with applications ranging from cancer to mobile health. Through these experiences, I realized that I enjoyed working closely with clinical collaborators to answer current scientific questions. I also realized that I was interested not only in the statistics, but also in understanding the scientific context behind the statistical work. This led me to my current position at CBAR. At CBAR, I work on clinical trials for the treatment and prevention of HIV and tuberculosis, collaborating closely with clinicians as part of the International Maternal Pediatric Adolescent AIDS Clinical Trials (IMPAACT) Network. My responsibilities range from developing safety monitoring and statistical analysis plans to reviewing case report forms and enrollment materials, as well as pursuing statistical methods research motivated by the clinical trial data. In this talk, I will describe my path to this position, my current responsibilities, and my experiences as an early-career collaborative biostatistician.

Training in a Data Coordinating Center: Fostering the Development of Highly Collaborative and Effective Early Career Biostatisticians

Megan McCabe, University of Alabama at Birmingham

Collaborative clinical research environments, and data coordinating centers (DCCs) in particular, require a diverse, interdisciplinary skillset, which is generally not the focus of graduate-level statistics or biostatistics coursework. Instead, these skills are typically learned and developed through research assistantships, internships, or other types of experiential learning opportunities.

Dr. McCabe trained in an academic DCC, the CTSDMC, and has pursued a career in academic DCCs post-graduation as the Assistant Director for Clinical Trials Development for the DATA CRU (UAB). Experiences as a graduate research assistant in a DCC were foundational for understanding the dynamic research environment associated with complex clinical research studies. At the CTSDMC, the speaker was the inaugural Network for Excellence in Neuroscience Clinical Trials (NeuroNEXT) Biostatistics Fellow and had the opportunity to gain experience working on clinical research studies from pre-award grant submission through dissemination and publication of results. In addition to academic DCC experience, she also completed two fellowships associated with FDA, as well as a co-op at a pharmaceutical company, Boehringer Ingelheim. This presentation will highlight the speaker's firsthand experience with training in a DCC environment and other experiential learning opportunities during her doctoral studies and how these experiences have facilitated her transition to faculty post-graduation.

Growing as a Collaborative Biostatistician in a Data Coordinating Center

Margaret Banker, Northwestern University

As an early-career faculty biostatistician in academic clinical research, I work in a Data Coordinating Centers (DCCs) where I collaborate with investigators to design, conduct, and analyze multicenter observational studies. This statistical collaboration involves developing statistical analysis plans, monitoring data quality, managing study data, and performing analyses that support study objectives and publications. In this talk, I will discuss experiences from two national research networks: the Glycemic Observation and Metabolic Outcomes in Mothers and Offspring (GOMOMS) study, coordinated through Northwestern University, and the Cure Glomerulonephropathy (CureGN) study, coordinated through the University of Michigan. I will highlight the diverse roles of biostatisticians in Data Coordinating Centers (DCCs) and offer an early-career perspective on the opportunities and challenges of collaborative biomedical research.

My Journey Through Clinical Trials: Experiences as a Faculty Member Working with a Data Coordinating Center

Emily Roberts, University of Iowa

As an Assistant Professor of Biostatistics at the University of Iowa, a significant proportion of my role is within our Clinical Trials Statistical and Data Management Center (CTSDMC) Data Coordinating Center (DCC). In particular, I work within the NeuroNEXT network to help design and analyze neurology-related clinical trials. I will share my experiences to date of transitioning from a graduate student to faculty member including my responsibilities within NeuroNEXT where I currently lead our pre-award team of biostatisticians.



Session Info

S301

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 00:30 - 1:00 UTC

Why and When Targeted Validation Sampling Offers Statistical Efficiency Gains: A Case Study on Healthy Food Access and Disease Outcomes

USE SCHED TO ACCESS THE ZOOM ROOM

Quantifying neighborhood food environments and understanding their relationships with residents' health is a public health priority. Using simple, error-prone food access metrics (like the shortest straight-line routes to healthy food stores) introduces measurement error and biases downstream statistical models, but measuring the more-accurate, map-based ones (like shortest driving routes) for entire studies is often implausible. Fortunately, adopting a two-phase design can harness the best of both metrics by combining the error-prone access measures for the entire study and the more-accurate ones for a chosen subset in a partial validation study. This validated subset can be strategically chosen to not only reduce bias but further improve efficiency when modeling relationships between health and the food environment. Technically, any information that is fully available for all neighborhoods can guide the validation sampling strategy. One such promising design paired stratification with Neyman allocation and sampled based on the influence function within each stratum. Using simulations and data for the Piedmont Triad Region of North Carolina, various validation sampling designs were evaluated to quantify the associations of diabetes count and obesity prevalence with neighborhood-level access to healthy foods, fitting two separate Poisson regression models, one for each outcome. We assess which design suits each model, and whether any is robust across settings.

ORGANIZER: L. Ishara Wijayaratne, Department of Statistics, Wake Forest University

CHAIR: Valeria Leiva, Universidad de los Andes



Speaker Bios

S301



MS. L. ISHARA WIJAYARATNE

Wake Forest University

<https://www.linkedin.com/in/ishara99/>

I am a graduate student in Statistics at Wake Forest University. Originally from Sri Lanka, I am keen on giving back to the community that has helped change my life so considerably. I am passionate about biostatistics, and my abstract submission addresses a pressing public health priority: expanding access to healthy food stores as a foundation for healthier living. As a woman in statistics, I am thrilled to take part in the IDWSDS conference and to contribute to a growing community of women advancing the field.



Session Info

S302

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 3:00 - 4:00 UTC

Emerging Computational Methods in Statistical Inference: Multivariate Analysis, Bayesian Shrinkage, and Advanced Regression Modeling

USE SCHED TO ACCESS THE ZOOM ROOM

This session presents novel computational and methodological advancements in modern statistical inference, focusing on complex data structures such as multivariate, longitudinal, and high-dimensional survival data. The presentations span both frequentist and Bayesian paradigms, featuring innovative strategies to overcome long-standing computational bottlenecks commonly encountered in such investigations. The session opens with a new framework for multivariate order-restricted statistical inference under a general grass ordering, applying a pairwise optimization criterion and introducing an efficient vector projection-based algorithm to estimate the order-restricted mean matrix. It then transitions into the Bayesian realm with two complementary presentations on longitudinal quantile regression. These back-to-back talks first present a robust partially collapsed Gibbs sampler to conquer convergence failures in high dimensions and then showcase a Bayesian adaptive LASSO framework that utilizes clustering-based shrinkage for simultaneous optimal quantile estimation and sparse variable selection. The session concludes by extending advanced continuous shrinkage priors to survival analysis, presenting a hybrid minorize-maximize algorithm for high-dimensional proportional hazards models using the horseshoe+ prior. Together, these talks highlight recent developments of highly efficient algorithms to extract reliable inferences from high-dimensional and complicated datasets.

ORGANIZER: Mai Dao, Wichita State University**CHAIR:** Mai Dao, Wichita State University

Speaker Bios

S302



MS. YUJIE JIA
Wichita State University

Yujie Jia is affiliated with Wichita State University, where she is currently a Ph.D. student in the Mathematics, Statistics, and Physics Department. Her research interests include multivariate order-restricted statistical inference, hypothesis testing, maximum likelihood estimation, design of experiments (ANOVA, MANOVA), and Bayesian quantile regression.



MS. SINING ZHANG
Wichita State University

Sining Zhang is a Ph.D. candidate in Applied Statistics at the Department of Mathematics, Statistics, and Physics at Wichita State University under the supervision of Dr. Mai Dao. Her research focuses on Bayesian quantile regression for longitudinal data, with an emphasis on partially collapsed Gibbs sampling, Markov chain Monte Carlo efficiency, and variable selection for both continuous and binary outcomes.



MS. SARAH GHAZAWNEH
Wichita State University

Dr. Sarah Ghazawneh is a Visiting Assistant Professor at the University of Connecticut. She earned her Ph.D. in Applied Mathematics from Wichita State University under the supervision of Dr. Mai Dao. Her research focuses on Bayesian quantile regression for longitudinal data, with an emphasis on efficient algorithms for parameter estimation and variable selection.



DR. MAI DAO
Wichita State University

Dr. Mai Dao is an Assistant Professor of Statistics in the Department of Mathematics, Statistics, and Physics at Wichita State University. She obtained her Ph.D. from Texas Tech University under the supervision of Professors Min Wang (The University of Texas at San Antonio) and Souparno Ghosh (University of Nebraska-Lincoln). Her primary research interests focus on Bayesian statistics, high-dimensional inference, and statistical machine learning.

https://www.wichita.edu/profiles/academics/fairmount_college_of_liberal_arts_and_sciences/Math/Faculty/Dao-Mai.php



Multivariate Grass Ordering and Pairwise Optimization Criterion*Yujie Jia, Wichita State University*

Multivariate order relations are generated from closed convex cones, the order-generating cones. Matrices with columns in orderings are confined to closed convex cones, the order-restricted cones. In MANOVA, the restricted MLE of the mean matrix in an ordering is the projection of the unrestricted MLE onto the order-restricted cone. This talk introduces a general grass ordering. By applying a pairwise optimization criterion summarized from cases where the restriction is only on two columns, a dynamic subset of the order-restricted cone for grass ordering is obtained. Based on the study of the matrix projections onto this subset, the projection of MLE onto the subset is proposed to be used as an estimator for the order-restricted mean matrix. This estimator is shown to be strongly consistent. Making use of this estimator in testing the homogeneity of the mean matrix leads to an ad hoc test. The test statistic of the test is shown to have a unique null distribution. The algorithm developed for the matrix projection onto the subset converts the problem of matrix projections to that of vector projections and can produce results in finite steps. A simulation study on the errors of the proposed estimator is conducted. The Monte Carlo method for computing the α -level critical values for the test statistic is discussed and illustrated.

A Partially Collapsed Gibbs Sampler for Bayesian Longitudinal Quantile Regression*Sining Zhang, Wichita State University*

Bayesian quantile regression for longitudinal data typically relies on the asymmetric Laplace distribution (ALD) working likelihood combined with Gibbs sampling via its normal-exponential mixture representation. However, the standard Gibbs sampler can suffer from slow mixing and poor convergence, particularly when the within-subject sample size or random-effect variance is large. We propose a Partially Collapsed Gibbs (PCG) sampler that replaces the standard full conditional with the marginalized conditional. A key ingredient is a ξ -reparameterization that decouples the scale parameter from the augmented-data covariance, yielding a closed-form inverse-gamma full conditional for the marginalized step. In an extensive simulation study, the proposed sampler achieves substantial gains in effective sample size per second over the standard Gibbs sampler, with the advantage growing sharply as within-subject sample size and variance increase. In several configurations, the standard Gibbs sampler exhibits catastrophic convergence failure, whereas the PCG sampler remains stable. The methodology is illustrated with an application to CD4 cell count data from an HIV/AIDS clinical study.

Bayesian Adaptive LASSO for Simultaneous Quantile Estimation and Variable Selection in Longitudinal Data*Sarah Ghazawneh, Wichita State University*

Quantile regression offers a flexible way to study covariate effects across the full distribution of a response, but in longitudinal and high-dimensional settings, estimating the quantile level and selecting relevant covariates at the same time

remains a challenge. We propose a Bayesian adaptive LASSO (BAL) framework that jointly estimates the quantile parameter τ and performs variable selection within a single model, built on the asymmetric Laplace mixture representation for Bayesian quantile regression. To address variable selection in high-dimensional settings, we incorporate a clustering-based shrinkage approach that groups regression coefficients according to their estimated effect sizes, allowing the model to adaptively distinguish between relevant and negligible covariates. We evaluate the method through simulation studies, examining its variable selection performance relative to fixed quantile alternatives. We also apply the method to real longitudinal datasets, including a low-dimensional economic application and a high-dimensional genomic application. We discuss how jointly estimating the quantile level and selecting variables through this clustering-based approach can improve both estimation accuracy and variable selection in longitudinal data analysis.

High-dimensional Proportional Hazards Model with the Horseshoe+ Prior*Mai Dao, Wichita State University*

In this talk, we discuss the use of horseshoe+ prior to the high-dimensional proportional hazards (PH) model for analyzing right-censored data. This work represents the first attempt to apply a continuous shrinkage prior in the setting of high-dimensional survival models and to compute the maximum a posteriori (MAP) estimator. Since the objective function is highly nonconcave, direct optimization of the posterior is infeasible. To address this challenge, we propose a hybrid minorize-maximize (MM) algorithm for optimizing the model parameters, which are partitioned into blocks containing nonoverlapping parameters. This decomposition, together with the careful construction of the minorizing function and the componentwise log-concavity of the objective function, greatly simplifies the maximization problem. Simulation studies demonstrate the effectiveness of the proposed algorithm, with favorable comparisons to several popular methods for high-dimensional survival models. Moreover, a near-optimal posterior contraction rate for the high-dimensional PH model is established, and a convergence rate for the MAP estimator is provided. Finally, we illustrate the practical utility of the proposed approach by identifying significant genetic biomarkers in an ovarian cancer study.



Session Info



CONTRIBUTED SESSION

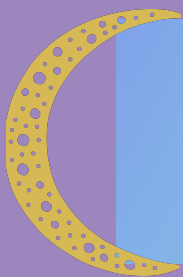
October 6th, 7:00 - 8:00 UTC

AI and Data Science 2

USE SCHED TO ACCESS THE ZOOM ROOM

Observation-Weight-Based Strategies for Robust Stochastic Gradient Boosting in Genomic Prediction | A journey of pivoting: From Local Research Experiences to Global Data Science Connections | Teaching Statistics in the Age of AI: Beyond Calculation in Health Sciences Education

S303

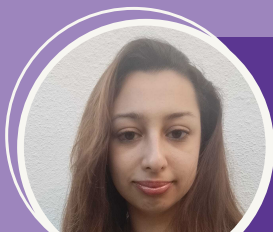


ORGANIZER: IDWSDS Program Committee

CHAIR: Malvika Kharbanda, SAiGENCI

Speaker Bios

S303

**MS. BEATRIZ GIL COMPARADO**

FCTNOVA Math

Beatriz Gil Comparado is pursuing a Master's degree in Mathematics and Applications – Specialization in Data and Decision Sciences at NOVA School of Science and Technology (NOVA FCT), Portugal. Her research focuses on statistical methodology and machine learning, with particular interest in robust predictive models for genomic prediction. Her current work investigates observation-weight-based approaches to improve the robustness of stochastic gradient boosting methods.

**MS. CHERLYNN DUMBURA**

CeSHHAR

<https://www.linkedin.com/in/cherlynn-dumbura/>

Cherlynn Dumbura is a Zimbabwean data scientist and health researcher working across climate-health, maternal and newborn health, statistics and research. She is a board member of Applied Malaria Modeling network and in the leadership of International Biometric Society, Zimbabwean regional chapter.

**PROF. CAMILA BERTINI MARTINS**

Federal University of São Paulo

Camila Bertini Martins is an Associate Professor at the Federal University of São Paulo (UNIFESP), Brazil, where she teaches Biostatistics in health sciences programs. She holds a PhD in Statistics from the University of São Paulo (USP). Her research interests include meta-analysis, Bayesian inference, survival analysis, and applied statistics. She also works on statistical education and innovation in health sciences teaching.



Observation-Weight-Based Strategies for Robust Stochastic Gradient Boosting in Genomic Prediction

Beatriz Gil Comparado, FCTNOVA Math

The presence of data contamination, such as errors or outliers, can severely compromise statistical models, including ensemble-based machine learning methods such as Stochastic Gradient Boosting (SGB), particularly in high-dimensional genomic prediction and selection studies. Although contamination may occur at both the response and covariate levels, this work focuses on response contamination and evaluates the robustness of the classical SGB method via simulation. Using a synthetic animal breeding dataset from the literature, we consider several plausible contamination scenarios and investigate robust adaptations of SGB based on observation-weighting strategies designed to enhance model resilience. Our study highlights the impact of data contamination on genomic prediction and explores robust alternatives to classical SGB that are suitable for routine use in breeding applications.

A journey of pivoting: From Local Research Experiences to Global Data Science Connections

Cherlynn Dumbura, CeSHHAR

Early career researchers are increasingly building careers across disciplines, sectors, and global research networks. For many, the challenge is not lack of skill but knowing how to translate existing training into a new professional direction. This talk designed for ECRs who are considering a pivot into data science, climate and health research, research operations, or scientific leadership, and who want to reposition themselves with clarity and confidence.

Using my journey as a Zimbabwean data scientist and public health researcher, I reflect on my journey from Biochemistry to Biostatistics to Data Science and beyond. I reflect on the process of building a career at the intersection of statistics, maternal and newborn health, climate change and research coordination in a cross sectoral consortium.

I reflect on how transferable skills from each point in my journey including leadership roles in scientific communities have shaped me in terms of strategy and scientific communication.

The talk offers a practical framework for career transition: Identifying an intellectual anchor, auditing transferable skills, selecting a technical edge, building visible proof of work, and using professional networks to access new opportunities. I also reflect on my own next phase of growth, including the integration of geospatial analysis, artificial intelligence, and climate and health decision-support tools into my research portfolio.

Teaching Statistics in the Age of AI: Beyond Calculation in Health Sciences Education

Camila Martins, Federal University of São Paulo

Statistics plays a central role in scientific knowledge production and evidence-based decision-making. In health sciences education, statistical literacy is essential for interpreting research findings, critically appraising the literature, and supporting clinical decisions.

Despite its relevance, teaching Statistics to health sciences students remains challenging. Many students experience anxiety toward discipline and struggle to connect statistical concepts with professional practice, often reducing learning to the memorization of formulas and statistical tests rather than meaningful reasoning.

The expansion of generative AI has added new complexity to this scenario. Statistical analyses, code generation, and result interpretation can now be partially automated, raising an important question: what should remain at the core of Statistics education?

This presentation discusses Statistics education in undergraduate health sciences programs considering these transformations. It argues that teaching should move beyond procedural execution and prioritize statistical reasoning, critical thinking, and data literacy. Special emphasis is placed on competencies such as formulating relevant questions, assessing data quality, recognizing methodological limitations, and critically interpreting evidence under uncertainty.

In the age of AI, Statistics education should focus less on calculations and more on preparing students to think critically about data, uncertainty, and evidence.



Session Info

S304

CONTRIBUTED SESSION

October 6th, 8:00 - 9:00 UTC

Education and Leadership 1

USE SCHED TO ACCESS THE ZOOM ROOM

A Bayesian satellite turbidity calibration algorithm framework with a focus on the Great Barrier Reef, Australia | A Statistical Framework for Educational Analytics Using Learning Management System Data: A Data-Driven Approach to Improving Student Learning | Implementing the ADAPT Model: Measuring Fidelity and Translating Research into Instructor Support | Implementation of Machine Learning Techniques for analyzing Autism Spectrum Disorder children through Art Emotion Extraction

ORGANIZER: IDWSDS Program Committee**CHAIR:** Vitalis Adero, AMREC

Speaker Bios

S304



NASIM SADRA
Massey University

I am a PhD candidate in New Zealand with a background in Civil Engineering. I later moved into remote sensing and data science, developing an interest in using these tools to better understand environmental processes. My current research focuses on satellite remote sensing and statistical modelling of coastal turbidity, particularly during extreme events. I am particularly interested in interdisciplinary approaches that bring together engineering, environmental science, remote sensing, and data science.



MS. ABINAYA NAGAMUTHU
Senior Lecturer

I am a Senior Lecturer in Statistics at IBSUniversity in Papua New Guinea and an Adjunct Lecturer at Southern Cross University with a strong academic background in statistics, business analytics, and econometrics. I hold an M.Phil. in Statistics and have completed a professional certification in Data Science and Data Analytics from Intellipaat. My teaching and research focus on applying statistical methods to solve real-world business and educational problems. I have experience teaching students in units/subjects such as Statistical Analysis, Business Mathematics, Business Analytics, and Quantitative Methods in Economics. I am committed to creating an engaging learning environment through face-to-face teaching, interactive classroom activities, Excel and R programming, case studies, quizzes, and Learning Management System (LMS)-based learning. My research interests include: Educational Analytics using LMS, Econometric modelling for Student academic Performance, etc.



MISS LAURIE SHORT
North Carolina State University

<https://mathedlaurieshort.wordpress.ncsu.edu/>

Laurie Short is a third-year PhD student in the Department of STEM Education at North Carolina State University. Her research focuses on college-level mathematics education, with a specific interest in the preparation of mathematics graduate teaching assistants. She has also researched discourse and social interaction in elementary mathematics, quantitative instruments for mathematics education, and spatial reasoning in upper-level math courses. Laurie serves as a research assistant for the Data Science and AI Academy, where she is studying the implementation of the All-Campus Data Science and AI Project-Based Teaching and Learning (ADAPT) model's 10 common learning elements.



PROF. C. VIJAYALAKSHMI
Central University of Tamil Nadu, Thiruvavur

<https://cutn.ac.in/staff-members/dr-c-vijayalakshmi/>

Professor & Head in the Department of Statistics and Applied Mathematics, Central University of Tamil Nadu. The teaching milestone records 25 years and 2 years industry. Published 136 research articles in journals with high IF and 79 papers in conference proceedings. The Google scholar citation is 320 with h-index 10 & i10-index 11. Published 5 patents and 5 monographs. She is also a recipient of 10 best paper awards. Completed the project in BRFSST for Rs. 30 lakhs, Seed Fund for Rs. 3 lakhs and as Project Director received a grant of Rs. 1.5 Crores from ICSSR. Guided 16 Ph.D. scholars and organized nearly 95 events at National and international levels. Recipient of many awards Best Professor by ASDF, Pondicherry Government, Lion's Club International, Raman & Ramanujam and Bharath Jyothi, MHRD-ICC, CUTN and Indo-Korea Science and Technology Centre innovation awards. With great zeal she is marching forward with determination to serve the student community and society at large.



A Bayesian satellite turbidity calibration algorithm framework with a focus on the Great Barrier Reef, Australia

Nasim Sadra, Massey University

Monitoring extreme events that are key drivers of coastal hazards such as flood related turbidity pulses and episodic water quality degradation depends critically on the statistical frameworks we use to retrieve them. Optical remote sensing-based turbidity algorithms typically rely on single spectral bands, band ratios, or index-based regression methods optimised for overall point accuracy. These approaches are primarily calibrated to the low turbidity conditions that dominate the available observations while extreme turbidity pulses critical for coastal hazard assessment, sediment transport dynamics, and infrastructure design are often underrepresented and underestimated. Using Sentinel 2 reflectance data (Level 2) matched with in-situ turbidity data from the Great Barrier Reef (Australia), this study develops a Bayesian regression framework based on maximum band ratio (MBR) and a well-established turbidity index called normalised difference turbidity index (NDTI). The model performance is assessed using leave one out cross validation (LOOCV), focusing specifically on prediction at extreme turbidity values rather than point accuracy alone. Preliminary results indicate that baseline model underrepresents uncertainty at high turbidity despite a good overall fit. This finding has direct implications for how remote sensing can be used for hazard and relevant extreme event modelling. The model structure will be extended to a hierarchical structure to test whether group specific structure improves calibration based on water turbidity regimes. By modelling the data generating process rather than predicting the existing values alone, this work aims to reduce overfitting and offer a transferable framework for uncertainty assessment in other remote sensing applications where rare events are disproportionately important while the dataset is dominated by low values. By doing so, this work aims to support long term coastal monitoring using freely available optical datasets and characterize relevant extremes that shape coastal resilience.

A Statistical Framework for Educational Analytics Using Learning Management System Data: A Data-Driven Approach to Improving Student Learning

Abinaya Nagamuthu, Senior Lecturer

In modern higher education, Learning Management Systems (LMS) have become an important platform for delivering teaching and learning activities. The use of LMS provides large volumes of digital learning data.

This talk presents a statistical framework for educational analytics that supports academics in making data-driven decisions in teaching environments. The LMS data includes attendance records, access to lecture materials, time spent on the LMS, completion of interactive activities, assessment scores, and final examination results. Descriptive statistics are used to summarise learning patterns, while correlation and regression analysis are applied to examine relationships between student engagement and academic performance. Predictive modelling techniques are also introduced to identify at-risk students and forecast their academic outcomes.

The proposed framework demonstrates how raw LMS data can be transformed into meaningful insights for academic decision-making. It supports improvements in course design, teaching strategies, and early intervention for at-risk students, ultimately enhancing student success rates. This talk also highlights the importance of integrating statistical methods with educational data analytics to promote evidence-based teaching practices in higher education.

Implementing the ADAPT Model: Measuring Fidelity and Translating Research into Instructor Support

Laura Short, North Carolina State University

This study investigates the implementation fidelity of the ADAPT model's 10 common learning elements across core data science curricula at North Carolina State University's Data Science and AI Academy. Utilizing Century, Rudnick, and Freeman's (2010) Critical Component Framework, the curriculum was evaluated across four foundational dimensions: Procedural, Educative, Pedagogical, and Student Engagement. To establish a comprehensive, inclusive data set, a qualitative case study was conducted across 7 distinct ADAPT courses. Entire semesters of instructional recordings, materials, and syllabi were reviewed to capture both explicit and subtle integration of the model's core elements in classroom settings. The findings indicate robust execution across several dimensions; however, four specific learning elements demonstrated a need for enhanced instructional support. These underrepresented elements include: explaining the scope of data science and AI expertise, introducing diverse career role models, articulating open research questions, and highlighting exciting discoveries within the discipline. To bridge these pedagogical gaps, empirical findings were directly translated into actionable faculty resources. High-fidelity classroom interactions identified during the data analysis were synthesized into practical, illustrative examples. This presentation demonstrates how implementation data can be leveraged to design targeted instructor support.

Implementation of Machine Learning Techniques for analyzing Autism Spectrum Disorder children through Art Emotion Extraction

Vijayalakshmi C, Professor, Department of Statistics and Applied Mathematics, Central University of Tamil Nadu, Thiruvavur

Autism spectrum disorder (ASD) mainly comprises of Neuro-degenerative disorders associated with social, communication, and behavioral difficulties. This paper mainly deals with the implementation of deep learning techniques to analyses and extract the emotions of autism children through their art using Deep Learning techniques. Feature extraction for the effective recognition of emotions is being obtained. It also targets on identifying the emotions through the artwork of children diagnosed with autism spectrum disorder (ASD) by implementing the predictions of their emotions which will help the doctors and caretakers so that the children will be in safe hands. This leads to the result of contributions to the state-of-the-art results in this research field. It also helps to analyses the ASD children's key emotions and the perceptions enabling them to analyses their emotional state.



Session Info

S305

TECHNICAL STATISTICAL SESSION

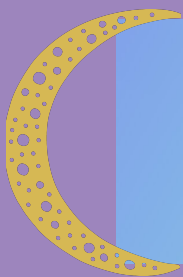
INVITED SESSION

October 6th, 9:00 - 10:00 UTC

The Role of Data Modelling for Better Outcomes in Women's Health

USE SCHED TO ACCESS THE ZOOM ROOM

Dr. Rachael Duncan will introduce the session that brings together leading speakers to explore how advanced data modelling can improve outcomes in women's health, particularly through equity-focused and locally led approaches. The session also highlights the importance of women as data scientists and modellers to ensure better healthcare outcomes. Dr Halima Twabi highlights capacity building in Sub-Saharan Africa, emphasising strengthening local expertise, supporting women data scientists, and applying advanced statistical and causal inference methods to complex health challenges. Dr Annabel Sowemimo contributes a critical perspective on structural inequalities, addressing racism and the need to decolonise healthcare systems. Dr Lucy Teece focuses on with insights from industry on the role of biometrics in cancer drug development. Together, the speakers demonstrate how inclusive, collaborative modelling approaches can generate more relevant, impactful, and equitable evidence for women's health research and policy.

ORGANIZER: Esther Jones, Biomathematics & Statistics Scotland**CHAIR:** Rachael Duncan, Biomathematics & Statistics Scotland

Speaker Bios

S305



DR. RACHAEL DUNCAN
BioSS

<https://www.bioss.ac.uk/people/rduncan>

Rachael Duncan is a Statistician in Animal Health and Welfare at BioSS. Since joining BioSS in 2023, she has applied statistical methods to a wide range of challenges in animal health and welfare. Prior to this, she completed her PhD at Lancaster University, where she explored how modelled data can be used to fill gaps in observational datasets and investigated how established methods for analysing large datasets perform in spatial applications. Her research interests centre on missing data and understanding the processes that generate observed data.



DR. HALIMA TWABI
University of Malawi

<https://twas.org/directory/twabi-halima-sumaya>

Halima Twabi is an Associate Professor of Statistics at the Department of Mathematical Sciences, University of Malawi. Her research interests are in causal inference for observational data, multivariate statistics, longitudinal and survival analysis, and statistical modelling and analysis of social and health data. In 2020, she received the L'Oréal-UNESCO for Women in Science Sub-Saharan Africa Young Talent Award for her work on applying causal inference to maternal and child health. In 2023, she was nominated treasurer of the International Biometric Society Sub-Saharan Africa Network, and treasurer of the IBS Malawi region. She is a founding member of the Girls in Science, Technology, Engineering, Arts and Mathematics network in Malawi; collaborates on "Strengthening Modelling and Analytics Capacity and Ecosystem for Women's Health" funded by the Gates Foundation, and the EU Sustainable Training in Epidemiological Analysis and Modelling for Data-Driven Decision Making (STEAM-3D).



DR. ANNABEL SOWEMIMO
King's College London

<https://www.annabelsowemimo.com/>

Dr Annabel Sowemimo is a doctor, academic, activist and writer, and an NHS Consultant in Sexual & Reproductive Health in South London. Her first book, *Divided: Racism, Medicine and Why We Need to Decolonise Healthcare* (2023), won the Bread and Roses Award for Radical Publishing and a Breast Cancer Book of the Year award, and was shortlisted for several major prizes. She founded the Reproductive Justice Initiative (formerly Decolonising Contraception), which aims to address health inequalities and racial disparities, and has received multiple awards for this work, including an outstanding achievement award from the British Association of Sexual Health and HIV (BASHH). Her interests include tackling Gender Based Violence and improving healthcare access for marginalised groups. Annabel is also a PhD candidate and Harold Moody Scholar at King's College London, researching the experiences of Black women in Britain with long-acting reversible contraception (LARCs).



DR. LUCY TEECE
AstraZeneca

<https://bsky.app/profile/stateecetics.bsky.social>

Dr Teece is an Associate Director of Statistical Science at AstraZeneca. Previously, she was a Lecturer in Medical Statistics in the Department of Health Sciences at the University of Leicester. She was awarded her PhD on "Investigating the presence and impact of competing events in prognostic model research" from Keele University in 2019. Her research interests focus on adapting statistical analysis methods to reflect the more complex realities observed in practice. Lucy has considerable experience in advanced survival analysis methods for developing and validating prognostic models, and linking and analysing large electronic health records data. During the pandemic she worked on projects to identify ethnic health inequalities in UK healthcare workers, and for her MRC secondment she is investigating the long-term effects of COVID-19 infection on cardiovascular, diabetes, and autoimmune health outcomes.





Introducing the Role of Data Modelling for Better Outcomes in Women's Health

Rachael Duncan, BioSS

Introduction to the session and the importance of including women statisticians and data scientists for better health outcomes.

Putting Women Data Modellers at the Centre of Women's Health Research in Sub-Saharan Africa

Halima Twabi, University of Malawi

When we talk about women's health in Sub-Saharan Africa, the focus is often on outcomes such as maternal mortality, HIV, and non-communicable diseases. Yet we rarely ask: Who is analysing the data, building the models, and interpreting results that shape policy decisions? Despite increasingly rich datasets, advanced analyses are often led externally, leaving local researchers, especially women, underrepresented. The SWRNAC project responds by placing women data modellers at the centre of capacity strengthening. This talk delves into the work the SWRNAC project is doing by equipping women researchers with advanced technical and analytical skills while promoting research through a gender lens. The project helps in asking how inequalities shape exposure, access, and outcomes. In doing so, it supports more inclusive, contextually grounded, and policy-relevant evidence for improving women's health.

Divided: Racism, Medicine and Why We Need to Decolonise Healthcare

Annabel Sowemimo, King's College London

Women in Biometrics: The Role of Statistics in Breast Cancer Drug Development

Lucy Teece, AstraZeneca



Session Info

S306

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 10:00 - 11:00 UTC

Modern Statistical Methods and Applications in Health Research

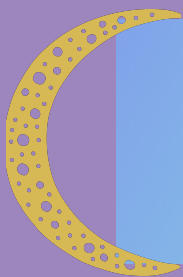
USE SCHED TO ACCESS THE ZOOM ROOM

This session highlights the breadth of biostatistics through four distinct but complementary talks by statistical researchers. It explores how modern statistical methodologies are advancing health research by addressing challenges in evidence generation, clinical decision-making, and causal inference. The presentations cover Bayesian learning frameworks for individualised treatment recommendations, large language model-assisted tools that improve the efficiency and reproducibility of systematic reviews and meta-analyses, innovative clinical trial designs and the practical challenges that influence their efficiency, and Mendelian randomisation approaches for investigating causal relationships and therapeutic targets in cardiovascular disease. Together, these talks aim to strengthen the reliability, transparency and translation of evidence, thereby supporting more informed clinical and public health decision-making in an increasingly connected global health research landscape.

ORGANIZER: Rajenki Das, MRC Biostatistics Unit, University of Cambridge

CHAIR: Rajenki Das, MRC Biostatistics Unit, University of Cambridge

SPONSOR: MRC Biostatistics Unit, University of Cambridge



Speaker Bios

S306



DR. PAMELA MASSIEL CHIROQUE SOLANO
University of Regensburg

Pamela Solano is a statistician passionate about transforming complex data into trustworthy clinical decisions. As a researcher at the Faculty of Informatics and Data Science, University of Regensburg, Germany, she develops interpretable AI and Bayesian learning methods that empower clinicians with transparent, evidence-based decision-support tools for personalized medicine. Her work integrates diverse real-world data to address complex, heterogeneous challenges across healthcare, ecology, actuarial science, epidemiology, and social media.



MS. DANYANG DAI
University of Queensland

Danyang Dai (Daidai) is a final year PhD student (thesis submitted) at the Queensland Digital Health Centre. She is an applied statistician and infectious disease epidemiology researcher with expertise in large scale health data analysis, disease surveillance, and advanced statistical modelling. Her research spans multinational COVID 19 studies, health economic evaluation, and the development of reproducible analytic pipelines. She is particularly interested in translating complex data into actionable insights, using robust methods to support evidence based public health policy, improve population health outcomes, and inform decision making in rapidly evolving health crises.



DR. ARITRA MUKHERJEE
Newcastle University

Aritra Mukherjee is a Research Associate in the Biostatistics Research Group at Newcastle University. Her research focuses on developing innovative statistical methods for clinical trials, with expertise in adaptive trial designs such as multi-arm multi-stage (MAMS) trials, sample size re-estimation and platform trials. She is passionate about advancing statistical methodology to make clinical research more efficient, reliable, and impactful. Through collaborations with clinicians and researchers, her work supports evidence-based healthcare and the translation of methodological innovation into real-world medical practice.



MS. DAIANNA GONZALEZ PADILLA
University of Cambridge

Daianna Gonzalez-Padilla is a researcher in Genetic Epidemiology at the MRC Biostatistics Unit, University of Cambridge, specializing in Mendelian randomization (MR), a causal inference framework that uses genetic variants to assess the effects of putative risk factors on disease from observational data. Her research integrates statistical evidence with molecular insights to identify plausible instrumental variables for MR and applies these frameworks to investigate the causal role of circulating biomolecules in cardiovascular disease across population strata, aiming to identify effective therapeutic targets and the subgroups most likely to benefit.



Bayesian Learning for Individualized Treatment Decisions in Personalized Medicine

Pamela Massiel Chiroque Solano, University of Regensburg

Personalized medicine seeks to identify the most appropriate treatment for an individual patient rather than relying solely on average treatment effects observed in clinical trials. However, translating heterogeneous patient information into reliable treatment recommendations remains a major statistical challenge.

This work presents a Bayesian statistical framework for estimating Predicted Individual Treatment Effects (PITE) by integrating evidence from clinical data. The approach captures complex treatment-response patterns while maintaining interpretability, allowing clinicians to understand how patient characteristics contribute to treatment recommendations.

Unlike conventional predictive models that focus primarily on prediction accuracy, the proposed framework explicitly supports treatment decision-making by combining individualized treatment effect estimation with decision-theoretic principles. It integrates high-dimensional and heterogeneous data while quantifying uncertainty through Bayesian Additive Regression Trees (BART). By producing transparent, patient-specific evidence, it supports clinicians in selecting treatments better aligned with individual patient profiles.

Applied to Alcohol Use Disorder (AUD), this work demonstrates how Bayesian statistical learning can inform individualized treatment decisions by integrating predictive modelling with principled decision support.

metaextractoR: An R Package for Data Extraction using LLMs for meta-analysis

Danyang Dai, University of Queensland

Systematic reviews and meta-analyses are central to evidence-based medicine, but data extraction is time-consuming and error-prone. Although Large Language Models (LLMs) show promise for automating this process, their adoption remains limited. We introduce metaextractoR, an open-source R package that streamlines data extraction by integrating open-source LLMs via the Ollama framework. The package supports automated, structured, and reproducible workflows through interactive Shiny applications, aligned with Cochrane guidelines and veridical data science principles.

metaextractoR includes three modular apps: (1) a manual extraction interface, (2) tools for prompt engineering and model selection, and (3) validation of LLM outputs. Together, these enable a human-in-the-loop approach, where each abstract is double-extracted by both a researcher and an LLM, mirroring best-practice recommendations. The package runs entirely on local machines, eliminating the need for APIs and ensuring data privacy. Comprehensive logging of prompts and outputs enhances transparency and reproducibility. By lowering technical barriers, metaextractoR enables more efficient, reliable, and transparent evidence synthesis.

Innovative trial designs: When not to use it

Aritra Mukherjee, Newcastle University

Over the last few years, the use of innovative trial designs has gained much momentum in practice. These designs encompass adaptive designs, platform trials, master protocols, among others. They offer greater flexibility without undermining trial integrity, often improving efficiency, patient benefit, and reducing the time to identify effective treatments. However, these advantages may be adversely affected by delays in observing the primary outcome variable. In the presence of such delays, the choice is between (a) pausing recruitment until sufficient data are accrued for the interim analysis, leading to longer times to identify effective treatment arms; or (b) continuing to recruit patients, which may result in many participants who do not benefit from the interim analysis. Little work has investigated the extent of outcome delay that results in the realised efficiency gains of these trial designs meeting the planned gains.

We assess the impact of outcome delay on the expected efficiency gains of different adaptive designs and platform trials by estimating the number of pipeline patients recruited under the assumption that recruitment is not paused while awaiting treatment outcomes.

The results indicate that if outcome delay is not considered at the planning stage, much of the expected efficiency gains may be lost. In light of these findings, we propose guidelines for clinicians when planning innovative trial designs in the presence of delayed outcomes.

Menopause-related heterogeneity in lipoprotein(a) effect on atherosclerotic cardiovascular disease using Mendelian Randomization

Daiana Gonzalez-Padilla, MRC Biostatistics Unit, University of Cambridge

Atherosclerotic cardiovascular disease (ASCVD) risk rises in women after menopause, making postmenopausal women a key prevention group. Lipoprotein(a) [Lp(a)] is a promising therapeutic target as it contributes to ASCVD risk and increases after menopause. However, whether its therapeutic potential could vary across menopausal stages remains unclear. Mendelian randomization (MR) uses genetic variants associated with an exposure to estimate its causal effect on an outcome. While MR is well suited to assess the role of Lp(a) in ASCVD, stratifying analyses by menopausal status may induce collider bias since menopause can be influenced by Lp(a)-related variants and ASCVD risk factors. In contrast, conditioning on sex and age avoids this bias, as neither is affected by Lp(a) genetics or cardiovascular risk factors.

Using the UK Biobank, we evaluated the effect of Lp(a) on coronary and peripheral artery disease across sex and age strata. Lp(a) was associated with increased risk in both women and men, with no evidence of female-specific divergence across age strata, supporting a consistent causal contribution of Lp(a) to ASCVD risk before and after menopause. More broadly, we propose sex- and age-stratified MR frameworks for studying menopause-related heterogeneity in cardiovascular aetiology.



Session Info

S307

OTHER

INVITED PANEL

October 6th, 12:00 - 13:00 UTC

Creating Alliances through a Data Feminism Community of Practice

USE SCHED TO ACCESS THE ZOOM ROOM

How has the Global Data Feminism Community of Practice (DFCoP) supported data practitioners and statisticians in leading inclusive practices in data collection and uptake? This question will guide the discussion in this session, which will feature 4 of the more than 150 members of the DFCoP. This session will introduce the DFCoP as a Global Majority international community, where practitioners, statisticians, researchers, and civil society from all around the world exchange experiences and knowledge on how to put data feminism into practice. The session will begin by contextualizing the community and speakers' strategic work in the fields of statistics and data feminism; as well as the practical ways in which members are co-creating a new network to continue the dialogue around feminist data practices and strengthen knowledge exchange. The session will then focus on the unique experiences of community members, including statisticians and project implementers. They will share some of their learnings and experiences on how they are applying data feminism into their statistical and data work. During the session, we will highlight the value of communities of practice like the DFCoP, and how members are not only practicing data feminism, but also collectively pushing the boundaries of this sector to advance equitable and accurate data practices. The session will conclude with a call to action, inviting attendees to join the community.

ORGANIZER: Karine Simonyan, GIZ Armenia**CHAIR:** Eleonora Sironi, Ladysmith Collective

Speaker Bios

S307



MRS. ANAHIT SIMONYAN
GIZ Armenia

Anahit Simonyan is a feminist researcher and human rights practitioner based in Armenia, with over 10 years of experience in combating gender-based violence and discrimination. Throughout her work she has cooperated with numerous local and international organizations and networks. Since 2018 she is also the Director of Human Rights Research Center (HRRC) - a non-governmental organization in Armenia. Since 2021 she is National Human Rights and Gender Expert at the Acopian Center for the Environment of the American University of Armenia (AUA). Anahit has authored and co-authored over 30 investigations, articles and publications that address, among else, issues around combatting GBV through legislation and policy, constitutional guarantees of human rights protection, the use of UN human rights procedure, protection of the rights of women human rights defenders and activists. Anahit holds an MA in Theory and Practice of Human Rights from the University of Essex, UK.



MRS. LUCINE KALANTARYAN
Statistical Committee of the Republic of Armenia (ARMSTAT)

<https://www.linkedin.com/in/lucine-kalantaryan-30147819b/>

Lucine Kalantaryan is the Head of the Labor Statistics Division and Gender Statistics Focal Point at the Statistical Committee of the Republic of Armenia (ARMSTAT), where she has coordinated the production and development of official statistics since 2002. An economist by training, she has over 25 years of experience in labor market statistics alongside more than a decade of expertise in gender statistics integrating data across sectors to measure gender inequalities and provide evidence that informs policy. She manages the implementation of multiple large-scale household and thematic surveys. Throughout her career, she has served as Deputy Project Director and Survey Methodologist on national and international initiatives, collaborating closely with the ILO, UN Women, the World Bank, GIZ, and the IOM. She has been a guest lecturer at the Armenian State University of Economics and contributes to international statistical capacity-building. Since 2025 Lucine is member of the DFCoP.



MISS ELEONORA SIRONI
Ladysmith Collective

<https://www.linkedin.com/in/eleonorasironi/>

Eleonora's research focuses on the intersections between gender, technology, and digitization. She brings cross-sectoral experience through work with NGOs like Doctors Without Borders and tech start-ups like VIOLA, an app focused on addressing gender-based violence. As a researcher at Ladysmith, they have been providing technical support for the global DFCoP since 2024 and co-designed the Data Feminism Capacity-Building Training aimed at Armenian data practitioners, including the Statistical Committee of the Republic of Armenia. Eleonora holds a MA in International Affairs, with a focus on International Security, from the Hertie School of Governance in Berlin and a BA in International Studies and European Institutions from the University of Milan.



DR. LORENA FUENTES
Ladysmith Collective

<https://www.linkedin.com/in/drlfuentes/>

Lorena's research explores how different types of data shape visibility and action in development, with a focus on gender-related violence. Currently, she is co-authoring a book about the 'gender data revolution' with her co-founder and feminist-thought-partner-in-crime, Dr. Tara Patricia Cookson. Lorena's work is informed by her previous experience as a frontline practitioner and lecturer at UCLA's International Institute. She earned her PhD from the University of London as a School of Social Sciences Fellow and Chair of the Gender & Social Policy in Latin America Academic Working Group. Lorena oversees Ladysmith's technology & digitization portfolio, among others, which currently includes the Data Feminism Community of Practice project.



Starting Off Strong*Nicole Dalzell, Wake Forest University*

First semester undergraduates are some of my favorite students to teach. They are excited, eager, and bring a bright energy into the classroom that no other group of students can replicate! Contrasting their excitement, however, is a struggle to adjust to the demands of college. This includes handling life responsibilities on their own, balancing academics and social goals, and dealing with stress and failure. For students with disabilities, this is also the first time students have had to advocate for their own accommodations. The process in K-12 is vastly different than it is in higher education, and requesting and implementing their accommodations college is a huge leap for these students. In this talk, we will discuss tips for structuring a course to support first year students in their transition to college. The emphasis on structure, clear expectations, and creating a safe space for failure and growth is designed to help students build a strong foundation for success in college regardless of their backgrounds in K-12. We will also chat about how some of this structure translates into courses for non-first year students to continue the process of helping students develop successfully in college.

A Focused First Impression*Ashley Mullan, Vanderbilt University*

In many K-12 environments, students gain little formal exposure to (bio)statistics and data science. Opportunities such as intensive summer courses can introduce students to ways that data interact with their daily lives. In this talk, we will discuss the development of a weeklong course geared towards high school students, incorporating the perspectives of both instructors and course alumni. We will cover the challenges of selecting and arranging course topics in an accessible order, creating engaging activities that break up the monotony of a full instructional day, and scaffolding a project that incorporates programming, analytical decision making, and communication. Attendees will leave with renewed energy for providing students with a focused first impression of (bio)statistics and data science.

Meeting Learners Where They Are: Teaching Statistics Across Diverse Audiences*Taylor Krajewski, Duke University*

Effective statistics education requires more than technical expertise; it requires understanding who the learners are, what motivates them, and how they will ultimately use statistical thinking in practice. Students enter quantitative learning environments, which can extend far beyond the traditional university lecture setting, with vastly different goals, backgrounds, and relationships with mathematics and statistics. Some aim to become statisticians or methodologists, while others seek to become informed collaborators, consumers of research, or practitioners working in increasingly data-driven fields. Drawing from experiences across traditional and nontraditional educational settings, including high school mathematics, graduate-level biostatistics instruction, interdisciplinary mentoring, and applied clinical research training, this talk will

reflect on lessons learned from tailoring statistical education for diverse audiences. The presentation will discuss approaches to adapting communication style, technical depth, examples, and instructional structure while maintaining rigor, engagement, and accessibility in a learning environment that builds both statistical reasoning and confidence. The talk will also reflect on how experiences across varied teaching and mentoring roles can teach educators to become more flexible, intentional, and effective communicators.

Tailoring Biostatistics Seminars for Emerging Investigators and Medical Trainees*Jamie Joseph, Henry Ford Health*

The United States Medical Licensing Examination (USMLE) Program is a series of three exams required for licensure to practice medicine in the United States. To become a practicing clinician, medical students and residents must pass all three exams, which include a biostatistics component. In this talk, we outline the process of designing and teaching a 20-part introductory biostatistics seminar series for residents and post-doctoral fellows. The series supports both USMLE preparation and skill development for trainees learning to lead scientific research projects independently. We will discuss strategies for tailoring lessons to a diverse audience with various levels of quantitative experience, recruiting appropriate speakers, and refining course content to ensure that students gain comfort with relevant epidemiological and biostatistical terminology and methods while also giving them the training for USMLE-specific examination questions.



Session Info



S308

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED SESSION

October 6th, 13:00 - 14:00 UTC

Pedagogy with Purpose: Tailoring your Statistics Classroom for Specific Audiences

USE SCHED TO ACCESS THE ZOOM ROOM

Given the wide variety of opportunities available for learners to gain expertise in statistics and data science, students enter the classroom with different needs and goals. Instructors balance standardized learning outcomes with individualized support, and a happy medium is not always easy to achieve. In this session, sponsored by the Caucus for Women in Statistics and Data Science, we feature a diverse group of educators whose trainees range from high school students just learning first principles to clinicians looking to build biostatistical expertise. The speakers will present their experiences customizing their courses to meet students where they are. First, Dr. Nicole Dalzell will share tips for structuring a first-year statistics course that builds a strong foundation for success in college. Next, Ms. Ashley Mullan will discuss lessons learned from designing a weeklong introductory summer course for high school students. Then, Dr. Taylor Krajewski will describe approaches for adapting communication and content while maintaining rigor, engagement, and accessibility. Finally, Dr. Jamie Joseph will outline considerations for adapting a biostatistics seminar series to provide support for both medical licensure candidates and emerging investigators. The four speakers span a wide array of educational settings, making the session interesting to a general audience. However, all four showcase how thoughtful education can support all members of the global statistics community.

ORGANIZER: Ashley Mullan, Vanderbilt University**CHAIR:** Jamie Roberman, Henry Ford Health

Speaker Bios

S308



DR. NICOLE DALZELL

Wake Forest University

Dr. Nicole Dalzell is an Associate Teaching Professor in the Department of Statistical Sciences at Wake Forest University. She loves developing new courses and constantly thinking about how to make her teaching materials and style work better for students. She also enjoys mentoring other faculty and collaborating with other educators in statistics education research. Outside of work, Nicole enjoys crocheting for family and friends, hiking, and pilates.



MS. ASHLEY MULLAN

Vanderbilt University

Ashley Mullan is a PhD student and research assistant in the Department of Biostatistics at Vanderbilt University. She earned her MS in Statistics from Wake Forest University. Her research interests include both methods development for measurement error and missing data and their applications to public health. As an educator, she showcases both her love of data and her commitment to advancing accessible scientific communication. In her free time, Ashley enjoys reading, crafting, and critically assessing her hobbies with a statistician's eye.



DR. TAYLOR KRAJEWSKI

Duke University

Dr. Taylor Krajewski is an Assistant Professor in the Department of Biostatistics & Bioinformatics and the Department of Population Health Sciences at the Duke University School of Medicine and a member of the Duke Clinical Research Institute. Her research focuses on causal inference, survival analysis, clinical trials, and the application of biostatistical methods to improve healthcare and public health decision-making. Dr. Krajewski spent several years as a high school mathematics teacher and department chair, experiences that continue to shape her approach to teaching, mentorship, and statistical communication. She has taught and mentored learners across a wide range of educational and professional settings, ranging from high school mathematics students to applied clinical collaborators. Dr. Krajewski is interested in tailoring statistical communication and instruction to diverse audiences while fostering confidence, engagement, and practical understanding of quantitative concepts.



DR. JAMIE JOSEPH

Henry Ford Health

Dr. Jamie Joseph is a faculty biostatistician in the Department of Public Health Sciences at Henry Ford Health with a joint appointment as an Assistant Professor of Research at Michigan State University. She earned her PhD in Biostatistics from Vanderbilt University in 2024. Currently, she collaborates with clinicians and researchers in all medical fields with particular focus on women's health, including maternal fetal medicine, endometriosis, fibroids, and urogynecology studies. Her methodological interests include causal inference, especially instrumental variable methods, applied to mobile health interventions.



Starting Off Strong*Nicole Dalzell, Wake Forest University*

First semester undergraduates are some of my favorite students to teach. They are excited, eager, and bring a bright energy into the classroom that no other group of students can replicate! Contrasting their excitement, however, is a struggle to adjust to the demands of college. This includes handling life responsibilities on their own, balancing academics and social goals, and dealing with stress and failure. For students with disabilities, this is also the first time students have had to advocate for their own accommodations. The process in K-12 is vastly different than it is in higher education, and requesting and implementing their accommodations college is a huge leap for these students. In this talk, we will discuss tips for structuring a course to support first year students in their transition to college. The emphasis on structure, clear expectations, and creating a safe space for failure and growth is designed to help students build a strong foundation for success in college regardless of their backgrounds in K-12. We will also chat about how some of this structure translates into courses for non-first year students to continue the process of helping students develop successfully in college.

A Focused First Impression*Ashley Mullan, Vanderbilt University*

In many K-12 environments, students gain little formal exposure to (bio)statistics and data science. Opportunities such as intensive summer courses can introduce students to ways that data interact with their daily lives. In this talk, we will discuss the development of a weeklong course geared towards high school students, incorporating the perspectives of both instructors and course alumni. We will cover the challenges of selecting and arranging course topics in an accessible order, creating engaging activities that break up the monotony of a full instructional day, and scaffolding a project that incorporates programming, analytical decision making, and communication. Attendees will leave with renewed energy for providing students with a focused first impression of (bio)statistics and data science.

Meeting Learners Where They Are: Teaching Statistics Across Diverse Audiences*Taylor Krajewski, Duke University*

Effective statistics education requires more than technical expertise; it requires understanding who the learners are, what motivates them, and how they will ultimately use statistical thinking in practice. Students enter quantitative learning environments, which can extend far beyond the traditional university lecture setting, with vastly different goals, backgrounds, and relationships with mathematics and statistics. Some aim to become statisticians or methodologists, while others seek to become informed collaborators, consumers of research, or practitioners working in increasingly data-driven fields. Drawing from experiences across traditional and nontraditional educational settings, including high school mathematics, graduate-level biostatistics instruction, interdisciplinary mentoring, and applied clinical research training, this talk will

reflect on lessons learned from tailoring statistical education for diverse audiences. The presentation will discuss approaches to adapting communication style, technical depth, examples, and instructional structure while maintaining rigor, engagement, and accessibility in a learning environment that builds both statistical reasoning and confidence. The talk will also reflect on how experiences across varied teaching and mentoring roles can teach educators to become more flexible, intentional, and effective communicators.

Tailoring Biostatistics Seminars for Emerging Investigators and Medical Trainees*Jamie Joseph, Henry Ford Health*

The United States Medical Licensing Examination (USMLE) Program is a series of three exams required for licensure to practice medicine in the United States. To become a practicing clinician, medical students and residents must pass all three exams, which include a biostatistics component. In this talk, we outline the process of designing and teaching a 20-part introductory biostatistics seminar series for residents and post-doctoral fellows. The series supports both USMLE preparation and skill development for trainees learning to lead scientific research projects independently. We will discuss strategies for tailoring lessons to a diverse audience with various levels of quantitative experience, recruiting appropriate speakers, and refining course content to ensure that students gain comfort with relevant epidemiological and biostatistical terminology and methods while also giving them the training for USMLE-specific examination questions.



Session Info



S310

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED SESSION

October 6th, 15:00 - 16:00 UTC

Read to Lead and Succeed in Statistics and Data Science

USE SCHED TO ACCESS THE ZOOM ROOM

In this session, you will learn why it is important for statisticians and data scientists to read books on leadership. You will appreciate how leadership books can inspire statisticians to be better leaders and collaborators. You will be introduced to the perspectives of statisticians about leadership. And you will be challenged to join or start a leadership book club to accelerate your career development.

ORGANIZER: Amanda Golbeck, UAMS**CHAIR:** Blagica Novkovska, University of Skopje

Speaker Bios

S310



DR. KSENIJA DUMICIC
University of Zagreb

Ksenija Dumičić, Ph.D., is a Full Professor Tenured in Statistics at the University of Zagreb Faculty of Economics & Business and a Visiting Professor at DOBA University of Applied Sciences in Slovenia. She founded and led Croatia's first postgraduate study programme in statistics for 15 years and served two terms as Head of the Statistics Department. She co-founded Ipsos Croatia in 1993 and the Croatian Statistical Association, serving as founding president and Honorary President. She chaired the ISI Committee on Women in Statistics for two terms, initiated the CW-ISI International Webinar Series, and received the 2025 ISI Service Award for exceptional leadership, dedication to ISI and the global statistical community; and her pivotal role in supporting and chairing CW-ISI. In 2026, she received the CWS 2025 Acknowledgement Certificate for service on the Country Representative Committee. She is a Fellow of the RSS and a member of the ASA and has published more than 260 papers.



DR. JAMES COCHRAN
College of Business, The University of Alabama

Dr. Cochran is Professor of Applied Statistics and the Mike and Kathy Mouron Research Chair with The University of Alabama. He was one of the founders of Statistics without Borders.



DR. JESSICA UTTS
University of California, Irvine

Jessica Utts is Professor Emerita of Statistics at the University of California, Irvine, where she served as department chair for 5 years. She has held numerous leadership roles in statistics organizations including as President of ASA, the Caucus for Women in Statistics, and WNAR; Chair of the Committee of Presidents of Statistical Societies (COPSS); Vice President of the International Statistical Institute (ISI) and of the International Association for Statistics Education (IASE), and numerous others. Before moving to UC Irvine, she spent three decades at UC Davis, where she served a term as the Associate Vice Provost for University Outreach.



DR. AMANDA GOLBECK
University of Arkansas for Medical Sciences (UAMS)

Amanda L. Golbeck is a statistician, academic leader, historian, biographer, and communicator of science. She is known for her authored books, "Equivalence: Elizabeth L. Scott at Berkeley" (Chapman & Hall, 2017) and "Florence Nightingale David: A Passionate Probabilist, Statistician, Historian, and Leader" (Springer, 2025). She is also known for her edited books on leadership, "Leadership and Women in Statistics" (Chapman & Hall, 2015, with Ingram Olkin and Yulia Gel) and "Leadership in Statistics and Data Science: Planning for Inclusive Excellence" (Springer, 2021). These are the only biographies of women statisticians in the 20th Century and the only books on leadership in statistics. Dr. Golbeck is a recipient of the Elizabeth L. Scott Award given by the Committee of Presidents of Statistical Societies. She is chair of the International Statistical Institute Committee on Women in Statistics.





Reading Leadership Books, Leading with Courage

Ksenija Dumicic, University of Zagreb, Faculty of Economics & Business, Croatia

This talk reflects on how leadership reading can strengthen women statisticians in a time of rapid change. Drawing on my own experience in academia, professional associations, project work, and mentoring, I show how leadership often begins in challenging or unexpected situations. I discuss lessons from selected books by Brené Brown, Sally Helgesen and Marshall Goldsmith, and Amanda L. Golbeck and colleagues, focusing on courage, authenticity, inclusivity, and continuous growth. I argue that technical expertise alone is not enough: effective leadership in statistics also requires empathy, communication, integrity, and the willingness to support others. Leadership books can help women statisticians build confidence, broaden their perspective, and lead with both data and courage.

Two Books that Inspire: "Eisenhower on Leadership" and "Mandela: Long Walk to Freedom"

James Cochran, College of Business, The University of Alabama

In this talk, we discuss leadership lessons that can be learned from these books and the lives of Nelson Mandela and Dwight D. Eisenhower.

Why Skillful Statistical Collaborators Make Skillful Leaders

Jessica Utts, University of California, Irvine

If you are a statistician and want to become an effective leader (formal or informal), you will find a wealth of ideas in the book "Leadership and Women in Statistics" (2016), edited by Amanda Golbeck, Ingram Olkin and Yulia Gel. The book contains essays from dozens of statisticians and leadership experts, covering a multitude of ideas, examples, and personal histories. In my talk I will present some of the ideas and stories I found most compelling. In the second part of my talk, I will discuss my views on leading an academic department, based on my own departments and on serving as an external reviewer for about a dozen statistics and math departments.

Developing a Leadership Book Club

Amanda Golbeck, College of Public Health, University of Arkansas for Medical Sciences

Leadership book clubs are an effective way to develop your power skills as you progress in your career. They can be designed to transform yourself, your team, your organizational unit, or your whole organization. By sustained silent reading and reflecting, people learn about leadership independently. By getting together with book club members to discuss and share reflections, people learn about leadership from each other. Shared reflections help club members to gain new perspectives, deepen connections with others, and accelerate learning. In this talk, I will outline the steps for building a leadership book club, and I will share my experience launching and leading a Leadership Classics Book Club for my university.



Session Info



S310

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED SESSION

October 6th, 15:00 - 16:00 UTC

Read to Lead and Succeed in Statistics and Data Science

USE SCHED TO ACCESS THE ZOOM ROOM

In this session, you will learn why it is important for statisticians and data scientists to read books on leadership. You will appreciate how leadership books can inspire statisticians to be better leaders and collaborators. You will be introduced to the perspectives of statisticians about leadership. And you will be challenged to join or start a leadership book club to accelerate your career development.

ORGANIZER: Amanda Golbeck, UAMS**CHAIR:** Blagica Novkovska, University of Skopje

Session Info

S311

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 17:00 - 18:00 UTC

Data Analysis in Engineering and Ecology: Reliability, Branching Processes and Random Fuzzy Sets

USE SCHED TO ACCESS THE ZOOM ROOM

This session brings together the last advances on Mathematical Statistics, and Data Analysis, considering the stochastic process framework, as well as more general and flexible probability modelling contexts, involving random elements, to represent uncertainty for data processing, in connection with Statistical Inference and Operations Research. Several practical problems can be analyzed, adopting the presented approaches, in Reliability, Engineering, Social Sciences, Ecology, and Medicine, just to mention a few.

Three invited speakers make up the invited session:

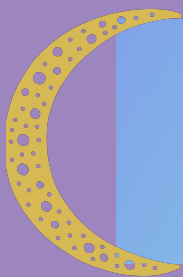
Lucía Bautista Bárcena, University of Extremadura (Spain).

Carmen Minuesa, University of Extremadura (Spain).

Beatriz Sinova Fernández, University of Oviedo (Spain).

ORGANIZER: María Dolores Ruiz-Medina, University of Granada

CHAIR: María Dolores Ruiz-Medina, University of Granada



Speaker Bios

S311



DR. LUCÍA BAUTISTA
Universidad de Extremadura

https://investigaex.com/socios/lucia_bautista_barcelona/

Lucía Bautista Bárcena is an Assistant Professor of Statistics and Operations Research at the University of Extremadura (Spain). She holds a BSc in Mathematics from the same institution (2017), where she also completed her MSc in Science Research (2019) and her PhD in Mathematics (2023). She was honoured with the YoungWomen4OR 2023 award by WISDOM (Women In Society: Doing Operational Research and Management Science) and EURO (The Association of European Operational Research Societies). Her research interests focus on the development of mathematical models applied to system degradation and maintenance optimization, combining tools from reliability, stochastic processes and statistical inference.



CARMEN MINUESA ABRIL
University of Extremadura

<https://sites.google.com/site/carmenminuesaabril/home>

Carmen Minuesa is an Associate Professor in the Department of Mathematics and a member of the Institute of Advanced Scientific Computing at the University of Extremadura (UEX). Her research focuses on stochastic processes. She began her career with a PhD on branching processes and has since been continuously involved in national research projects. She has held a position as assistant professor at the Autonomous University of Madrid and completed several research visits at international universities. Her work has been recognized with the Juan Jesús Morales Scientific Research Prize (Faculty of Sciences, UEX) and the Ramiro Melendreras Prize (SEIO). She also represented Spain at the European Young Statisticians Meeting organized by the Bernoulli Society. She has served on scientific and organizing committees of international conferences and co-coordinated the SEIO working group on Stochastic Processes and Their Applications (2022–2025).



BEATRIZ SINOVA
University of Oviedo

<https://cientificasinnovadoras.fecyt.es/cientificas/beatriz-sinova-fernandez>

Beatriz Sinova Fernández is an Associate Professor in the Department of Statistics and Operational Research and Mathematics Didactics at the University of Oviedo. She holds a Bachelor's degree in Mathematics from the University of Oviedo, a Master's degree in Mathematical Modelling, Statistics and Computing, and a joint PhD in Mathematics and Statistics from the University of Oviedo and Ghent University (Belgium). Her research focuses on developing robust statistics techniques for imprecise-valued data, mainly interval- and fuzzy-valued data. Currently she leads the research group on "Statistical Methods with Imprecise Random Elements Comparison of Distributions of Random Elements" at the University of Oviedo. She has published more than 25 papers in JCR-indexed journals. Among the awards she has received are the Ramiro Melendreras Award from the SEIO and the Vicent Caselles Award from the Royal Spanish Mathematical Society and the BBVA Foundation.



Statistical inference for Wiener degradation models under imperfect maintenance and general observation schemes

Lucía Bautista, Universidad de Extremadura

Degradation modelling is essential in reliability engineering and predictive maintenance. In real-world industrial applications, systems frequently undergo imperfect maintenance that only partially restore their health, and degradation measurements are often irregular or incomplete. This work introduces a framework to overcome these practical limitations. We model the underlying continuous degradation using a Wiener process and represent the effect of imperfect maintenance through the Arithmetic Reduction of Degradation (ARD) model. We develop a Maximum Likelihood Estimation (MLE) methodology capable of handling a general observation scheme, including missing data and varying inspection schedules (before, after, or between maintenance actions). This model is extended to bivariate degradation processes, using a shared-noise approach to capture environmental dependencies between multiple components. To support practical decision-making, we provide prediction intervals that capture the uncertainty of both the model and the parameter estimation. The applicability of this methodology is demonstrated through a real-world case study on critical components from the energy sector, showing how this framework can effectively predict future degradation and optimize maintenance scheduling.

robust statistics, several alternatives for summarizing the central tendency (and others for the scale) of fuzzy-valued data have been introduced in the literature. In particular, trimmed means and location M-estimators have shown a good finite-sample performance under contamination, and their robustness has been proved by means of the finite sample breakdown point. These measures and more recent approaches, such as fuzzy medoids, will be illustrated using a real-life example.

Estimating the carrying capacity of endangered species

Carmen Minuesa, Universidad de Extremadura

In this talk we consider population-size-dependent branching processes (PSDBPs), which are stochastic models that describe the dynamics of populations living in habitats where the resources are limited and whose individuals reproduce independently of the others.

We focus on PSDBPs whose extinction occurs almost surely and, in a parametric framework for the offspring distribution, address the estimation of the parameters of the model. Based on the observation of the population sizes, a new least-square estimator is introduced. The asymptotic properties of this estimator are analysed and its C-consistency and asymptotic normality are established. The results are applied to a real dataset on a threatened bird: the black robin population.

This is a joint work with Peter Braunsteins (UNSW Sydney) and Sophie Hautphenne (University of Melbourne).

Robust estimation of central tendency for fuzzy-valued data

Beatriz Sinova, University of Oviedo

Imprecision is inherent to many real-life situations from fields as diverse as Social Sciences, Medicine or Engineering. A useful way to cope with it is to consider random fuzzy sets, for which an appropriate statistical methodology has been developed for decades. Nevertheless, most of these techniques involve non-robust central tendency measures, which may compromise conclusions when data contain outliers or errors. Inspired by



Session Info

S312

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 18:00 - 18:30 UTC

New advances on Functional data model-based clustering

USE SCHED TO ACCESS THE ZOOM ROOM

Functional data analysis has attracted considerable attention in recent years, and its applications appear in physical processes, genetics, biology, meteorology, and signal processing. Many modern applications produce data best viewed as functions rather than finite-dimensional vectors because of their nature. Beyond the challenges of collecting and preprocessing such data, efficiently handling large volumes of functional observations has become an urgent concern. On one hand, functional data takes values in an infinite-dimensional space, which is challenging to handle with classical methods. On the other hand, ignoring the functionality aspect of data will lead to information loss. In this talk, we highlight model-based clustering methods, a powerful tool in machine learning for identifying subgroup-specific patterns. Specifically, we introduce new model-based clustering techniques for functional data, regardless of the Gaussian assumption. The performance of each algorithm is evaluated through simulations and real-world datasets, and the results confirm their efficiency.

ORGANIZER: Mina Aminghafari, University of Calgary**CHAIR:** Candace Berrett, Brigham Young University

Speaker Bios

S312

**DR. MINA AMINGHAFARI**

University of Calgary

<https://profiles.ucalgary.ca/mina-aminghafari>

Dr. Mina Aminghafari is an Associate Professor in the Department of Mathematics and Statistics at the University of Calgary. Her research lies at the intersection of high-dimensional statistics, machine learning, and applied data science, particularly on clustering theory and statistical learning and their applications in health and environmental sciences. She has contributed significantly to methodological advances in regularization techniques, co-clustering, and functional data analysis. Her work frequently bridges theory and practice, motivated by real-world challenges in biomedical diagnostics, autoimmune disease research, and environmental monitoring.



Session Info

S313

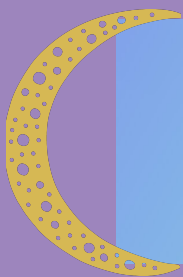
CONTRIBUTED SESSION

October 6th, 20:00 - 21:00 UTC

Education and Leadership 2

USE SCHED TO ACCESS THE ZOOM ROOM

Understanding the role of non-academic factors in student success: Perspectives from a minority-serving institution | Gender Norms and Economic Decision-Making: Evidence on Risk and Labor Outcomes | Faculty Teaching and Research Productivity and Equity in STEM: Student Grade Gaps in Lower-Division Courses | Elevating Collaborations with Leadership and Hospitality

ORGANIZER: IDWSDS Program Committee**CHAIR:** Tiffany Timbers, University of British Columbia

Speaker Bios

S313



DR. SINJINI MITRA

California State University, Fullerton

Dr. Sinjini Mitra is a professor in the Information Systems & Decision Sciences (ISDS) department at the College of Business and Economics, California State University, Fullerton (CSUF). Previously, she was a postdoctoral research fellow at USC's Information Sciences Institute. Dr. Mitra earned her Ph.D. in Statistics from Carnegie Mellon University. She is a faculty fellow of CSUF's Catalyst Center and Center for Cybersecurity and is the co-Director of the Center for Information Technology and Business Analytics. Her research interests encompass statistical applications in security, healthcare, business, and education, emphasizing interdisciplinary settings. Dr. Mitra has authored two books, several journal articles, and book chapters, and served as a Co-PI on two NIH collaborative grants focused on student mentoring and research. She also serves as the Director of Undergraduate Assessment at her college. Dr. Mitra is a member of IMS, ASA, and INFORMS.



MRS. KRISTIN GAFFNEY

University of Arkansas for Medical Sciences

Kristin Gaffney is a collaborative staff biostatistician in the Department of Biostatistics shared between the College of Medicine and the College of Public Health at the University of Arkansas for Medical Sciences. She has a Master of Public Health degree from the University of Nebraska Medical Center and a Bachelor of Arts degree in Mathematics from Bethel University, Minnesota. She has worked in academia for the last four years after spending over 20 years in the non-profit sector supporting marriages and families. Kristin enjoys helping others succeed and spending time with her husband and adult children.



MRS. ORLIE WEITZMAN

University of Chicago

I am a junior at the University of Chicago Laboratory Schools with an interest in behavioral economics, particularly gender and wealth inequality. I have worked as a student researcher for Professor Marianne Bertrand and am currently involved in an ongoing study on the global rise of ideological polarization between men and women. Outside of research, I am a student journalist whose writing has been syndicated internationally and has received a National American Voices Award from the Scholastic Art and Writing Awards. I am also a youth community advocate for immigration and migrant rights, and have organized state-level lobbying efforts on these issues.



Understanding the role of non-academic factors in student success: Perspectives from a minority-serving institution

Sinjini Mitra, California State University, Fullerton

Since the COVID-19 pandemic, much research has examined how educational institutions adapted to the abrupt shift to online learning, as well as the factors influencing student learning during that period and the challenges encountered by educators and administrators alike. One important aspect of student learning that was brought to the forefront by the pandemic was the significance of non-academic and social/environment factors in student performance and success that had received much less attention in prior studies compared to demographic and academic factors. This research study explores different non-academic factors that play a role in college success, particularly at a minority-serving university with a diverse student population, although many of the findings apply to most institutions today. Statistical analysis of the data revealed that factors such as students' self-perceived academic and emotional readiness for college, motivation and self-confidence, participation in campus co-curricular activities, perception of a sense of community and belonging, and the use of academic support services were significantly associated with student performance and success. Practical implications and future directions from both an academic and an institutional perspective will also be briefly discussed.

Gender Norms and Economic Decision-Making: Evidence on Risk and Labor Outcomes

Orlie Weitzman, University of Chicago

The research I supported examines the economic impact of gender identity and social norms, emphasizing the distinction between biological sex and the culturally constructed process of "doing gender" in economics. It argues that internalized beliefs and social expectations surrounding gender shape key life outcomes, including academic performance, labor force participation, and household specialization. The talk focuses on three empirical patterns: (i) gender norms and risk-taking, (ii) occupational sorting, and (iii) labor market outcomes across countries.

Within this framework, gender differences in economic behavior—particularly risk-taking—are shaped by internalized norms about how men and women are expected to behave. Risk-taking is identified as a core dimension of dominant masculinity norms, and cross-country evidence suggests that men's adherence to these norms is associated with greater willingness to take risks. Using data from more than 70 countries, De Haas et al. (2024) find that stronger conformity to masculinity norms is significantly associated with higher male risk tolerance, as well as broader patterns of occupational sorting and labor market behavior. Taken together in my co-report, these findings suggest that incorporating gender identity and norms into economic analysis meaningfully improves our ability to explain both individual behavior and aggregate labor market gender representation patterns.

Elevating Collaborations with Leadership and Hospitality

Kristin Gaffney, UAMS

To be a successful statistician or data scientist, we have to work skillfully with data. To be a successful collaborator, we also have to work skillfully with people. People are not data. While we focus on technical skills like coding, study design, statistical methods, and now AI, we should also refine skills that enable us to thrive in teamwork. Leveraging resources outside statistics and data science can equip us to be successful both technically and interpersonally. I will share lessons learned from leadership education and hospitality literature so you can take the next step in elevating your collaborative capabilities and experiences.



Session Info

S315

TECHNICAL STATISTICAL SESSION

PLENARY TALK

October 6th, 21:00 - 22:00 UTC

AI and Data Science 3

USE SCHED TO ACCESS THE ZOOM ROOM

From Law to Data Governance: Rebuilding from Scratch at the Intersection of Legal Frameworks and AI | AI in Long-Term Care: Practical Tools for Workflow Efficiency, Documentation Support, and Caregiver Well-Being | Which one to use, traditional Algorithms or AI? Approaches for Computational Problem-Solving | Optimizing Portfolio Allocation with Unsupervised Machine Learning

ORGANIZER: IDWSDS Program Committee

CHAIR: Kristin Gaffney, University of Arkansas for Medical Sciences



Speaker Bios

S315



MISS CHARITY NYAMUCHENGWA
Clarkson University

Charity is an applied AI practitioner focused on improving workflows, documentation, and safety in long-term care environments. She develops practical, caregiver-centered tools including an AI-assisted documentation helper, a medication-workflow guidance system, and a synthetic early-warning model designed to support staff in high-pressure care settings. Her work emphasizes operational efficiency, caregiver well-being, and the responsible use of AI in healthcare. Charity's current projects explore how lightweight AI systems can reduce administrative burden, strengthen communication, and enhance daily care routines in long-term care facilities.



DR. WEIJIE PANG
Wentworth Institute of Technology

Weijie Pang is a tenure-track assistant professor in School of Computing and Data Science at Wentworth Institute of Technology. After received Ph.D. from Worcester Polytechnic Institute and B.S. and M.S. from Beijing University of Technology, Dr. Pang finished the Post-doc program in Department of Mathematics and Statistics at McMaster University. Her research area is applied data science, financial mathematics, pandemic modeling and stochastic analysis.



MISS DANIELA MÁRQUEZ OLGÚN
Universidad Autónoma de Chiapas

<https://www.linkedin.com/in/daniela-m-009b7a35b/>

I am a lawyer-turned-data-scientist based in Aguascalientes, Mexico. After more than eight years working in Mexico's Federal Judiciary, I pursued an M.Sc. in Data Science, Big Data and Business Analytics at Universidad Complutense de Madrid and am currently completing a B.Sc. in Statistics and Information Systems. I recently worked as a Data Operations Intern at Bridge for Billions, supporting data governance and quality control for a global social impact programme. My interests include AI governance, data policy, and digital rights.



DR. JAE SOOK CHEONG
AMExNet

Dr. Jae Sook Cheong is a senior researcher and a data professional. Her works are in the area of data governance, knowledge representation with data analysis, algorithm analysis, and machine learning. With many years of experience across academia, industry, and national R&D institutes, her work bridges theoretical computer science and applied digital innovation. She has contributed to small and large-scale R&D projects in research data management, knowledge graphs and digital content technology. She also has supported interdisciplinary digital transformation initiatives in an international environment. Her current research interests focus on representing knowledge structure using AI model and data modelling using ontology. Her interest also includes the societal implications of data and artificial intelligence. She earned her Ph.D. from Utrecht University and an M.Phil. from HKUST. They are both in theoretical computer science. She holds a B.A. in computer science from POSTECH.



From Law to Data Governance: Rebuilding from Scratch at the Intersection of Legal Frameworks and AI

Daniela Márquez Olguín, Universidad Autónoma de Chiapas

It started with two papers.

During my Master's in Constitutional Justice, I encountered for the first time research that genuinely combined legal analysis with statistical thinking. Then, already working inside the Federal Judiciary of Mexico, I read a study on nepotism in the Mexican judicial system that used three different statistical strategies to measure the same concept. As a lawyer, something immediately felt off: before measuring, you have to define. Those two moments planted a question legal training alone couldn't answer: how do you actually measure what's happening inside an institution you know from the inside?

That question led me toward data science and data governance — and to a harder truth: my last math class had been in high school, over a decade earlier.

This talk is an honest account of that pivot: nearly seven years of hesitation shaped by the belief that choosing a career is a permanent, irreversible decision; the process of rebuilding algebra, functions, and statistics from scratch — calculus being, unexpectedly, the hardest part; and the leap of leaving a stable judicial career for a full-time role in data operations, doubts included.

The slow discovery was this: a non-technical background is not a liability. It is a perspective. Someone who spent eight years inside an institution knows something that no dataset alone can tell you — that before you measure something, you have to define what it means. That is exactly what AI governance needs and rar

AI in Long-Term Care: Practical Tools for Workflow Efficiency, Documentation Support, and Caregiver Well-Being

Charity Nyamuchengwa, clarkson university

This talk shows how practical AI tools can reduce documentation load, streamline caregiver workflows, and improve daily operations in long-term care facilities. Using lightweight, privacy-safe systems including a caregiver-support chatbot, workflow guidance tools, and synthetic early-warning models we explore how AI can boost efficiency and support caregiver well-being without relying on patient-specific data. Attendees will leave with a clear understanding of how applied AI can make long-term care safer, smoother, and more sustainable.

Which one to use, traditional Algorithms or AI? Approaches for Computational Problem-Solving

Jae Sook Cheong, AMExNet

There are many technologies available to solve problems, such as AI, traditional exact algorithms, probabilistic approaches, and many more. To apply the right technology effectively, we must understand which types of problems each technology is best suited for. In this talk, we will explore different computing models

and problem space, and see how AI approaches differ from other algorithms and approaches. This will help us determine when to use AI and when other approaches are more appropriate.

Optimizing Portfolio Allocation with Unsupervised Machine Learning

Weijie Pang, Wentworth Institute of Technology

Unsupervised learning offers a data-driven pathway to portfolio construction without relying on strong parametric assumptions. In this talk, I present a framework that leverages three clustering techniques—k-means, self-organizing maps (SOM), and fuzzy c-means—to group securities based on shared risk and performance characteristics. From each cluster, representative assets are selected to form diversified portfolios using three simple selection rules: highest average closing price, highest return, and lowest volatility.

I evaluate the resulting portfolios through cumulative return analysis and benchmark their performance against the S&P 500 and the 10-Year Treasury yield. To examine robustness, I further conduct stress tests during two major market disruptions: the 2008 financial crisis and the 2020 COVID-19 pandemic.

The comparative analysis highlights how different clustering methods influence portfolio composition and performance. Overall, the results demonstrate the potential of unsupervised machine learning to enhance diversification and deliver competitive performance using only publicly available data. This work provides practical insights into integrating clustering techniques into systematic portfolio design.



Session Info

S316

CONTRIBUTED SESSION

October 6th, 22:00 - 23:00 UTC

AI and Data Science 4

USE SCHED TO ACCESS THE ZOOM ROOM

Bayesian Time-Dynamic Mixed-Effects Modeling of Longitudinal Data | When the Data Goes Quiet: Using HDBSCAN to Reveal Gendered Street Crime Patterns Hidden in Official Crime Data | From Gut to Data: A Replicable Forecasting Framework for Rural Micro-Retailers in Data-Poor Environments | Collective Intelligence Lead to Open Innovations

ORGANIZER: IDWSDS Program Committee**CHAIR:** Azizur Rahman, Government of Manitoba, Canada

Speaker Bios

S316



DR. NISHA DANIEL RAJU DANIEL
Bayer AG

<https://www.linkedin.com/in/drnisha-daniel-8ab49325/>

Dr. Nisha Daniel R is a qualified Dental Surgeon seamlessly bridging the worlds of clinical medicine and advanced digital health informatics. Currently completing her Master of Science in Data Science for Health and Social Care at the University of Edinburgh, her expertise lies at the intersection of healthcare, artificial intelligence, and clinical research informatics. Deeply driven by open innovation, she specializes in leveraging cutting-edge machine learning, Natural Language Processing (NLP), and speech signal analysis to solve complex clinical bottlenecks. Her current work focuses on precision psychiatry, the democratization of clinical domain expertise through conversational AI ("vibe coding"), and building scalable, patient-centric digital health applications.



MS. MAKENA GRIGSBY
University of California, Riverside

Makena Grigsby recently earned her M.S. in Statistics from the University of California, Riverside. Her academic work focuses on Bayesian modeling, survival analysis, machine learning, and statistical computing. She is especially interested in developing and applying flexible statistical models for complex data structures, including longitudinal, biomedical, transcriptomic, and ecological data.



MS. FARAH AHMED
HORUS VISTA LABS and OSU

Farah Ahmed is a Senior Analytics and Insights Lead at Equitable Advisors and the founder and Principal Data Scientist at Horus Vista Labs, an applied data science consultancy serving rural and underserved small businesses. Her work spans demand forecasting, causal inference, and behavioral analytics, with a focus on making machine learning interpretable and adoptable for stakeholders with no formal data infrastructure. Her anchor engagement with a SNAP-authorized rural grocery store in South Dakota, in partnership with Warehouses4Good, reduced forecasting error by over 40% using a stakeholder-validated modeling approach. She holds a Master of Data Analytics from The Ohio State University and has presented her work at ASA SDSS 2026 and the DSI Pedagogy Conference. Her broader research interests include spatial-temporal crime analysis and applied ML for underserved communities.



Bayesian Time-Dynamic Mixed-Effects Modeling of Longitudinal Data*Makena Grigsby, University of California, Riverside*

Longitudinal studies often collect repeated measurements on subjects over time, making it important to account for within-subject dependence and changing covariate effects. In many applications, measurements are not collected on a common schedule; subjects may have different numbers of observations, irregular follow-up times, or covariates that vary over time. This work presents a Bayesian time-dynamic mixed-effects model framework for longitudinal outcomes with balanced or irregular observation schedules. Time-varying regression coefficients are modeled using spline basis expansions, allowing covariate effects to change smoothly over the observed time domain while subject-specific random effects account for dependence among repeated measurements. Model parameters, including coefficient functions, variance components, and random effects, are estimated using Markov chain Monte Carlo sampling, with posterior summaries used for inference and uncertainty quantification. The framework is implemented in JAGS for Gaussian, Bernoulli, and Poisson outcomes, allowing the approach to be applied to continuous, binary, and count longitudinal data. Simulation studies examine performance under balanced longitudinal data, irregular measurement times, and time-varying covariates. Overall, this work provides a flexible Bayesian framework for modeling longitudinal data when covariate effects are not constant over time and observation schedules may vary across subjects.

When the Data Goes Quiet: Using HDBSCAN to Reveal Gendered Street Crime Patterns Hidden in Official Crime Data*Farah Ahmed, HORUS VISTA LABS and OSU*

Official crime taxonomies were not built to surface how gendered violence shapes daily life in public space. When sexual assault, stalking, and harassment are folded into broad categories dominated by property crime and theft, the patterns that determine how women move through cities disappear into statistical noise.

This talk presents an unsupervised learning approach applied to over 387,000 reported crimes in Chicago from 2013 to 2025, filtered to non domestic incidents involving gendered violence. Rather than relying on predefined categories or administrative boundaries, we used HDBSCAN, a density based clustering algorithm, to let spatial and temporal patterns emerge on their own. The model incorporated geographic coordinates alongside a composite risk score weighting factors such as location isolation and time of day.

The results showed that just over a third of gendered street crimes form dense, geographically stable clusters, meaning the same neighborhoods surface as high risk year over year. The remaining crimes appear as scattered incidents that traditional hotspot mapping would miss entirely, requiring different intervention strategies than concentrated crime does.

This methodology is not specific to Chicago or to any single reporting system. Every city that tracks crime using categories built around property offenses has the same blind spot: the crimes that most affect how women navigate public space are the

crimes most likely to vanish in the data.

From Gut to Data: A Replicable Forecasting Framework for Rural Micro-Retailers in Data-Poor Environments*Farah Ahmed, HORUS VISTA LABS and OSU*

Small independent retailers in rural and underserved communities often operate on intuition, not by choice, but because formal data infrastructure does not exist. There are no point-of-sale systems, no inventory software, and no historical transaction records, yet these stores are frequently the only food access point for their communities.

This talk presents a replicable three-layer data science framework developed with a SNAP-authorized independent grocery in rural South Dakota that had no existing data infrastructure. Starting from manually collected sales records, we built a demand forecasting pipeline that progressed through SARIMA, SARIMAX, XGBoost, and ultimately linear regression, chosen not for sophistication but because the store owner could validate its coefficients against decades of operational experience. Stakeholder adoption, rather than model complexity, defined success.

SHAP values applied during the XGBoost phase revealed which features actually drove demand and informed the decision to prioritize an interpretable final model. The framework's roadmap extends to inventory intelligence and a plain English LLM interface that would let owners query their data conversationally. The methodology was designed for portability: minimum viable data collection, interpretability over accuracy maximization, and owner validation as a feature rather than a limitation. This same framework could be applied to any small business or micro retail across the globe.

Collective Intelligence Lead to Open Innovations*Nisha Daniel Raju Daniel, Bayer AG*

Open innovation hackathons have evolved from isolated corporate experiments into an essential strategic framework within the pharmaceutical sector to drastically reduce research costs, accelerate clinical pipelines, and democratize cross-disciplinary domain expertise. This abstract demonstrates how global leaders like Bayer, Novartis, Pfizer, Johnson & Johnson, and Sanofi utilize high-energy "vibe coding" environments to bridge traditional science-tech divides, transforming passive internal workforces into active solution builders who crowdsource critical advancements in data science, trial optimization, rare disease diagnostics, and automated regulatory processing. While these dynamic sprints offer immense advantages by enabling rapid prototyping, modular tool creation, and immediate risk validation without heavy capital expenditure, they simultaneously present unique organizational vulnerabilities, including complex intellectual property licensing, internal resistance to change management, and long-term technical debt. Ultimately, this study redefines how the life sciences industry evaluates innovation success, shifting the metric from finalized software products to the velocity of insights generated and long-term behavioral transformation.



Session Info

S400

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

INVITED SESSION

October 6th, 8:00 - 9:00 UTC

Building Future Leaders in Statistics and Data Science Japan's Collaborative Model from School Education to Society

USE SCHED TO ACCESS THE ZOOM ROOM

As artificial intelligence and data-driven decision-making become integral to society, developing future leaders in statistics and data science has become an international priority. Beyond teaching statistical methods or programming skills, education must cultivate statistical thinking, inquiry, evidence-based reasoning, and the ability to connect data with real-world decision-making throughout learners' educational journeys.

Japan has been developing an inquiry-based learning framework that progressively connects statistical and data science education from primary school to university. This framework encourages learners to investigate authentic questions using data while strengthening statistical literacy through curriculum reform, inquiry-based learning, student competitions, and university education and research. It is supported through collaboration among educational institutions, professional societies, public organizations, and non-profit organizations.

This session introduces four complementary perspectives on this framework. First, it presents how long-term educational support by a non-profit organization has contributed to nurturing future statistical talent. Second, it explores how national and international statistical competitions, including the Statistical Graph Contest and the ISLP International Poster Competition, inspire students and foster statistical thinking. Third, it shares innovative practices in inquiry-based data science education in Japanese secondary schools. Finally, it discusses how statistics and data science education can prepare hybrid talent capable of integrating disciplinary expertise with AI and emerging technologies across secondary education, higher education, and lifelong learning.

Through these perspectives, the session aims to share Japan's inquiry-based educational framework with the international community and to promote dialogue on how statistics and data science education can cultivate future leaders capable of addressing increasingly complex societal challenges.

ORGANIZER:

CHAIR: Yutaka Kano, The Japan Statistical Society

SPONSOR: Sinfonica and Japan Statistical Society



Speaker Bios

S400



SHIGERU KAWASAKI
The Japan Statistical Association

Shigeru Kawasaki is the President of the Japan Statistical Association, a non-profit organization aiming to contribute to the development of statistics and dissemination of statistical knowledge to society at large.

He has a long career as a government statistician and a statistical educator. He was Director-General of the Statistics Bureau Japan in 2007-2011. After leaving the government, he was engaged in statistical research and education as Professor at Nihon University.

He has been actively engaged in international statistical activities. He served the International Association for Official Statistics as the President in 2013-2015. He also served the United Nations Statistical Commission as the Chair-man in 2011 and 2020-2021.



KIWA TOMARU
Nagoya University Junior and Senior High School

Kiwa Tomaru is a mathematics teacher at Nagoya University Junior and Senior High School, where she has designed and taught the compulsory Data Science course since its introduction in 2022, emphasizing reading data correctly over formula-based instruction — work recognized with the 19th Japan Statistical Society Award for Statistics Education. She is currently a Doctoral Program student in Degree Programs in Systems and Information Engineering at the University of Tsukuba, researching the relationship between regional amenities and well-being, bridging secondary-education practice and graduate research to make statistical thinking accessible to society.



YOKO KONISHI
University of Tsukuba

Yoko Konishi is a professor at the University of Tsukuba, specializing in applied econometrics and statistics. She received her Ph.D. in Economics from Nagoya University. Her research uses official statistics, private-sector big data, and large-scale surveys to study consumer behavior, disaster recovery, and the societal impact of generative AI.

She is also committed to promoting statistical literacy and making statistical thinking an essential part of education and decision-making in society.

Session Info

S401

TECHNICAL STATISTICAL SESSION

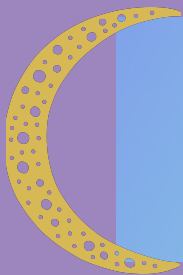
CONTRIBUTED SESSION

October 6th, 9:00 - 10:00 UTC

Health and Biostatistics 1

USE SCHED TO ACCESS THE ZOOM ROOM

A Background Correction and Normalisation Framework for multiplex
Immunofluorescence Spatial Proteomics | Menopause-related heterogeneity in lipoprotein
(a) effect on atherosclerotic cardiovascular disease: a sex- and age-stratified Mendelian
Randomization study in the UK Biobank | Income, Inequality, and Mortality: A
Harmonized Comparative Framework Across Ecuador, Colombia, and Brazil|Absorbing
Markov Chain Parameter Estimation Under Data Scarcity: A Comparative Study of
Analytical and Monte Carlo Methods in Neonatal Care|Predicting the Right Treatment for
the Right Patient: An AI-Powered Decision Support Framework Based on Predicted
Individual Treatment Effects

ORGANIZER: IDWSDS Program Committee**CHAIR:** Abisola Esther Abdulmalik, University of Medical Sciences,
Ondo

Speaker Bios

S401



MS. MALVIKA KHARBANDA
SAIGENCI

Malvika is a PhD candidate in computational oncology at Adelaide University (Australia), where she develops statistical and computational methods to analyse complex biomedical data. Her research focuses on using large scale clinical and spatial imaging datasets to better understand how diseases such as cancer develop and progress.

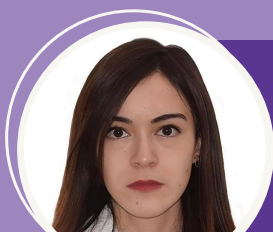
She is currently developing `bgnorm`, a statistical framework that uses probabilistic modelling to separate true biological signal from background noise in high dimensional protein imaging data. Her work aims to improve the reliability and interpretability of modern biomedical datasets where thousands to millions of measurements are collected across tissue samples.



DR. GABRIELA SANDOVAL
CISEAL-PUCE

<https://www.linkedin.com/in/gabysandovalm/>

PhD in Statistics, currently working as a postdoctoral researcher at the Center for Research on Health in Latin America (CISEAL), where she contributes to projects aimed at reducing health inequalities and mitigating the impact of economic crises in the region. Her expertise centers on the analysis of censuses, surveys, and administrative records, as well as the development of statistical models to study social determinants and health outcomes across Latin American countries. She also promotes the use of reproducible tools and data science to strengthen evidence-based decision-making.



MS. DAIANNA GONZALEZ PADILLA
University of Cambridge

Daianna Gonzalez-Padilla is a researcher in Genetic Epidemiology at the MRC Biostatistics Unit, University of Cambridge, specializing in Mendelian randomization (MR), a causal inference framework that uses genetic variants to assess the effects of putative risk factors on disease from observational data. Her research integrates statistical evidence with molecular insights to identify plausible instrumental variables for MR and applies these frameworks to investigate the causal role of circulating biomolecules in cardiovascular disease across population strata, aiming to identify effective therapeutic targets and the subgroups most likely to benefit.



MX. LEANDRA BRAEUNINGER
University College London

<https://leandrabraeuninger.github.io/>

Leandra Bräuninger (they/them; she/her) is a doctoral student at University College London, supervised by Dr Brieuc Lehmann and Prof. Ioanna Manolopoulou, using statistical and machine learning methods to define, quantify and mitigate genomic inequity. Drawing on algorithmic fairness, semi-parametric Statistics and Gaussian processes, conformal prediction, and applying it to topics such as equitable polygenic risk scores and variant effect prediction, Leandra's research leverages a variety of disciplines to solve complex problems at their intersection. Previously, Leandra held research and teaching positions in statistical fairness in genomics, mathematical malaria prediction, and fundamental biology at the Alan Turing Institute, the Mathematical Biology Group at the University of Melbourne, and the Centre for Active Learning at ETH Zurich. Leandra holds an MSc in Mathematical Modelling of Global Health (Oxford) and a BA in Data Science and International Relations (St. Gallen, CH).



A Background Correction and Normalisation Framework for multiplex Immunofluorescence Spatial Proteomics

Malvika Kharbanda, SAiGENCI

Multiplex spatial proteomics enables simultaneous imaging of dozens of proteins within intact tissue sections, generating detailed maps of protein expression across thousands of cells. Recent advances now allow these measurements to extend across multiple tissue slices, producing large three-dimensional datasets. However, the raw intensity values produced by these technologies contain substantial technical noise arising from instrument effects, tissue autofluorescence, and non-specific antibody binding. These background signals introduce bias and make it difficult to distinguish true biological signal from noise. Current approaches to background correction and quality control rely largely on manual inspection or hard thresholding, which are subjective, variable and impractical when analysing hundreds of proteins across hundreds of samples.

We present bgnorm, a model-based framework for pixel level background correction using a 3 component Gaussian mixture model. Pixel intensities are modelled as arising from background noise, non-specific signal, or true biological signal. Using this model, we estimate and remove background contributions through probabilistic signal deconvolution. The framework also includes model based quality control to detect failed stains and a normalisation step to enable comparisons across samples and channels.

bgnorm improves separation between signal and background, enabling more reliable downstream spatial analysis of high dimensional imaging data

Menopause-related heterogeneity in lipoprotein(a) effect on atherosclerotic cardiovascular disease: a sex- and age-stratified Mendelian Randomization study in the UK Biobank

Daiana Gonzalez-Padilla, University of Cambridge

Atherosclerotic cardiovascular disease (ASCVD) risk rises in women during the menopausal transition, making postmenopausal women a key target for prevention. Lipoprotein(a) [Lp(a)] is a promising therapeutic target as it contributes to ASCVD risk and increases after menopause. However, whether its therapeutic potential could vary across menopausal stages remains unclear.

Mendelian randomization (MR) uses genetic variants associated with an exposure to estimate its causal effect on an outcome. While MR is well suited to assess the role of Lp(a) in ASCVD, stratifying analyses by menopausal status may induce collider bias since menopause can be influenced by Lp(a)-related variants and ASCVD risk factors. In contrast, conditioning on sex and age avoids this bias, as neither is affected by Lp(a) genetics or cardiovascular risk factors. Sex- and age-stratified MR can thus be used to investigate menopause-related heterogeneity in Lp(a) effects, conceptualized as age-dependent differences specific to women.

Using the UK Biobank, we evaluated the effect of Lp(a) on

coronary and peripheral artery disease across sex and age strata. Lp(a) was associated with increased risk in both women and men, with no evidence of female-specific divergence across age strata. These findings support a consistent causal contribution of Lp(a) to ASCVD risk across the female life course and propose sex- and age-stratified MR frameworks for studying menopause-related heterogeneity in cardiovascular aetiology

Income, Inequality, and Mortality: A Harmonized Comparative Framework Across Ecuador, Colombia, and Brazil

Gabriela Sandoval, CISEAL-PUCE

Inconsistent income measures and non-comparable covariate structures across national census systems often limit comparative mortality research across Latin America. We present a harmonized analytical framework based on census microdata from Ecuador, Colombia, and Brazil, designed to enable direct cross-country comparisons of income-mortality gradients. The estimates reveal meaningful heterogeneity in the strength of the income-mortality relationship across countries. This talk offers both a reusable methodological pipeline for researchers working with limited-income-data census contexts and substantive findings on cross-national mortality inequality patterns

Towards continuous ancestry aware PRS via genetic distance kernel methods

Leandra Braeuninger, University College London

Polygenic risk scores suffer from well-documented portability failures across the genetic ancestry continuum, arising from several interacting sources: systematic differences in linkage disequilibrium patterns, misspecification of allele frequency normalisation, potential effect size heterogeneity, and GWAS discovery bias towards European-ancestry cohorts. To date, more than 87% of GWAS have been conducted in predominantly European cohorts, and the resulting PRS show substantially degraded performance in non-European individuals. Critically, this degradation is not confined to discrete ancestry group boundaries: Ding et al. demonstrated that PRS accuracy decays continuously with an individual's Euclidean distance from the training data centroid in PC space, establishing that portability is fundamentally a problem of genetic distance rather than of categorical ancestry misassignment. The prevailing methodological response of partitioning individuals into continental ancestry groups and estimating group-specific corrections temporarily improves performance within those strata but entrenches the conflation of biological and sociopolitical categories that the field has been called upon to abandon. A principled solution requires methods that operationalise ancestry as a continuous variable from the outset. Here we present a kernel formulation of PRS which subsumes previous continuous PRS methods and allows for kernel design choices that maximise portability directly.



Session Info

S402

TECHNICAL STATISTICAL SESSION

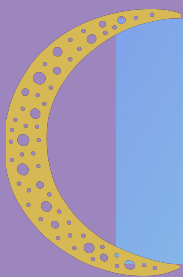
CONTRIBUTED SESSION

October 6th, 10:00 - 11:00 UTC

Selected Topics in Statistical Methods and Applications

USE SCHED TO ACCESS THE ZOOM ROOM

Beyond the Average: Random Slope Models for Revealing Territorial and Gender Inequalities | Randomization Inference on Policy Assignments | Becker's models for mixture experiments: An A-optimal approach | Inference for GMANOVA model for Volatile Data

ORGANIZER: IDWSDS Program Committee**CHAIR:** Marilena Müller, German Cancer Research Center

Speaker Bios

S402



DR. NAOMI DIZ ROSALES
Universidade da Coruña, CITIC

I hold a PhD in Statistics from the Universidade da Coruña (UDC), where my research focused on the development of mixed models with random slopes for Small Area Estimation. My work combines statistical methodology with applications to socioeconomic and public health challenges, including poverty analysis from a gender perspective and the prediction of healthcare resource demand. My academic background spans Biology, Bioinformatics, and Statistics, reflecting an interdisciplinary interest in applying data science to real-world problems. Throughout my research, I have worked on developing statistical methods that support evidence-based decision-making and help uncover inequalities that are often hidden by aggregated data. Alongside my research, I enjoy combining scientific work with university teaching and outreach activities to promote statistics and data science among new generations and to highlight their potential for addressing real-world societal challenges.



DR. EUNYI CHUNG
University of Illinois at Urbana-Champaign

She is an Associate Professor of Economics at the University of Illinois Urbana-Champaign, specializing in econometric theory. Her research focuses on developing improved methods of statistical inference for evaluating the causal effects of policy interventions using both experimental and observational data. Much of her recent research centers on permutation tests and other resampling-based methods for causal inference.



PROF. BUSHRA HUSAIN
Aligarh Muslim University, Aligarh, UP

<https://www.amu.ac.in/faculty/statistics-and-operations-research/bushra-husain>

Bushra Husain, Professor is the Chairperson of Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, U.P., India and has been working as Faculty of Statistics at Women's College since 2003. She has obtained B.Sc., M.Sc., M.Phil., and Ph.D. in Statistics from University of Delhi, Delhi. Her primary research area is Design of Experiments, and she specializes in Mixture Experiments and Optimization. She has participated in a number of National and International Seminars/Conferences organized both within and outside India and has co-authored twenty research papers published in journals of international repute. Current profile includes teaching and research guidance. Supervised one M. Phil and four Ph. D Scholars in the field of Design of Experiments. Life member of Indian Science Congress Association, Indian Society for Probability and Statistics, Indian Society of Agricultural Statistics and Society of Statistics, Computer and Applications.



MISS SUDEEPTA PAL
PhD Scholar at IIT, Hyderabad

I am currently pursuing my PhD in Statistics at IIT Hyderabad, where I also work as a Teaching Assistant. My research mainly focuses on statistical theory, multivariate analysis, and computational statistics. I cleared two of India's toughest competitive exams with high ranks, securing All India Rank 138 in GATE 2024 and AIR 174 in IIT-JAM 2022. Before coming to IIT Hyderabad, I completed my master's degree in Statistics from Utkal University, graduating second in my department with a 9.05 CGPA. For my bachelor's degree at Rama Devi Women's University, I graduated as the top student in my entire department. I also won a Nationwide Quiz Contest on Official Statistics held by the National Statistical Organisation (NSO). Along with my academic work, I have gained hands-on experience as a data science intern and recently served on the organizing committee for the CDSDM'26 conference at IIT, Hyderabad.



Beyond the Average: Random Slope Models for Revealing Territorial and Gender Inequalities

Naomi Diz Rosales, Universidade da Coruña, CITIC

Socioeconomic indicators based on regional averages often conceal important territorial and gender inequalities. This work shows how incorporating random slopes into Small Area Estimation models uncovers heterogeneous relationships between poverty, income, and their socioeconomic determinants across territories and population groups.

Using Spanish official data, the analyses reveal that the protective effect of higher education against poverty varies substantially across provinces and between men and women. Likewise, the impact of unemployment and demographic factors differs across income groups, reflecting distinct socioeconomic dynamics. A longitudinal analysis of disposable income further highlights regional convergence alongside persistent gender disparities in economic growth.

These findings demonstrate that random slope models provide more than improved statistical fit: they reveal meaningful socioeconomic heterogeneity that can support more effective, place-based, and gender-sensitive public policies.

Randomization Inference on Policy Assignments

EunYi Chung, University of Illinois at Urbana-Champaign

Placebo randomization tests are becoming a widely used statistical approach in empirical economics.

Common examples include difference-in-differences designs, where researchers randomize the labels of treatment and control groups, as well as regression discontinuity and bunching designs, where researchers randomize the discontinuity point. It is understood that this testing procedure has exact size control in finite samples under the sharp null hypothesis. However, size control requires researchers to know the distribution of the policy assignment, that is, the data generating process that assigns individuals into treatment and control groups. This paper proposes a new placebo randomization test that relaxes this requirement. Our test controls size in large samples under arbitrary misspecification of the policy distribution, while it controls size exactly in finite samples when the policy distribution is correctly specified. The paper makes two additional contributions. First, we demonstrate that in designs prone to biased test statistics --- such as regression discontinuity and bunching --- the size distortion of our test does not necessarily increase with the bias. Second, we address settings where statistics evaluated at all placebos are perfectly collinear and thus not exchangeable, a problem that arises in difference-in-differences. We show how to implement our test in these cases while maintaining its size control features.

Simulations show desirable finite sample.

Becker's models for mixture experiments: An A-optimal approach

Bushra Husain, Aligarh Muslim University, Aligarh, UP

In this article, we have investigated the derivation of weight expressions for the A-optimal designs for Becker's models using the weighted simplex-centroid design approach and established a comprehensive generalized form for these weights. Employing the equivalence theorem, we have validated the derived allocations as A-optimal designs. Furthermore, we have conducted a comparative efficiency analysis against the D-optimal design, shedding light on the performance disparities between the two methodologies.

Inference for GMANOVA model for Volatile Data

Sudeepta Pal, PhD Scholar at IIT, Hyderabad

The Generalized Multivariate Analysis of Variance (GMANOVA) model is widely used for analyzing longitudinal and growth curve data with structured mean profiles, where both the between (groups) and within individual (across time) design matrices are incorporated in a bilinear set up. The classical inference for the GMANOVA model is developed under the distributional assumption of matrix-variate normality which makes it sensitive to skewed and volatile data. The objective of the study is to propose a robust inference for the GMANOVA model for data exhibiting volatility such as financial data, assuming the error matrix follows a matrix variate t . The Maximum Likelihood estimation of the unknown regression coefficient matrix B and unknown variance covariance matrix σ is developed. Since the equations do not admit closed formed solutions, an iterative approach like Newton Raphson is carried out. To achieve this, the model is formulated via a hierarchical (latent variable structure) representation of the matrix variate t distribution. Three distinct iterative architectures are implemented and evaluated: the standard Newton Raphson method, the Expectation Maximization (EM) algorithm and the Expectation-Conditional Maximization Either (ECME) algorithm. The simulation performance of the proposed t -based iterative estimators demonstrates strong robust alternatives with significantly lower MSE and numerical stability. Future work includes developing hypothesis testing procedures.



Session Info

S403

TECHNICAL STATISTICAL SESSION

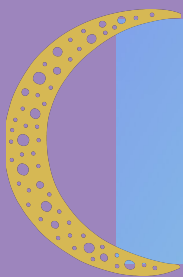
INVITED SESSION

October 6th, 12:00 - 13:00 UTC

Change Point Detection and Its Applications

USE SCHED TO ACCESS THE ZOOM ROOM

Change point detection has long been an active and important area of statistical research because structural changes frequently arise in data collected over time or across different conditions. Detecting these changes helps researchers better understand complex processes and make more informed decisions in fields such as public health, finance, genomics, environmental science, manufacturing, and engineering, where the ability to detect shifts in trends, variability, or dependence structures can provide valuable insights into complex systems. As the volume and complexity of modern data continue to grow, the development of robust and efficient change point techniques remains a significant research challenge and opportunity. This invited session brings together researchers from three countries to present new methods and real-world applications of change point analysis. The session will highlight recent advances in the field, encourage international collaboration, and showcase the broad impact of change point research across diverse disciplines.

ORGANIZER: Jie Chen, Augusta University**CHAIR:** Jie Chen, Augusta University

Speaker Bios

S403



DR. REBECCA KILICK
Clemson University

Rebecca Killick is Associate Director of Research in Mathematical and Statistical Sciences at Clemson University. Dr. Killick received a PhD in Statistics from Lancaster University, where until 2026, held a Professor and Director of Research positions. In 2019 Dr. Killick was the first UK recipient of the “Young Statistician of the Year” award from the European Network for Business and Industrial Statistics. Dr. Killick’s primary research interests lie in development of novel methodology for the analysis of univariate and multivariate nonstationary time series models and Rebecca is highly motivated by real world problems and has worked with data in a range of fields including Bioinformatics, Energy, Engineering, Environment, Finance, Health, Linguistics and Official Statistics. Rebecca is Editor-in-Chief of the Journal of Statistical and Graphical Statistics, and the Journal of Statistical Software, and a founding member of the ROpenSci statistical software peer review board.



DR. KUN LIU
JPMorgan Chase & Company

Kun Liu is a Vice President and Quant Modeling Lead at JPMorgan Chase with expertise in AI, machine learning, credit risk, fraud detection, and causal inference. He holds a Ph.D. in Statistics from Georgia Tech and has over eight years of experience developing and deploying advanced analytics solutions in financial services.



MS. MEENU RANI
Indian Institute of Technology Ropar

Meenu Rani is a Ph.D. student in the Department of Mathematics at the Indian Institute of Technology Ropar. Her doctoral research focuses on change point detection, including both offline and online settings. Her work has primarily involved applications to financial data.



DR. AYTEN YİĞİTER
Hacettepe University

Dr. Yiğiter received the B.S., M.S., and Ph.D. degrees in statistics from Hacettepe University, Ankara, Türkiye, in 1995, 2000 and 2006, respectively. In 2007, she was a Visiting Scholar with the University of Missouri-Kansas City, Kansas City, MO, USA. She is currently working as an Associate Professor with the Department of Statistics, Hacettepe University, Ankara, Türkiye. Her current research interests include change point and mixture models.



Developing changepoint methodology motivated by climate applications*Rebecca Killick, Clemenson University*

Global warming is always a hot topic and whatever side of the fence you sit on, you cannot deny the data that the climate is changing - regardless of attributing a cause to that. These changes have huge impacts on individual lives all over the world. In the UK we have had warmer summers and milder winters (at least in southern parts of the UK!) and more extreme weather events than earlier in the twentieth century. Most people have a storm, flooding or drought story, or know someone who has.

In this talk, I will share some of my own stories about working with the data behind the headlines, mostly in temperature time series from across the globe. I will provide an introduction to the changepoint models I use to describe these temperature data and provide commentaries on working with environmental scientists and media engagement. From the statistical side, I am hoping that you will take away some insights into why changepoint models are useful, an intuition of (and some R packages for) how to apply these models to any data, and the key challenges you may come across in doing so. In littering my talk with anecdotes about interdisciplinary work, I am hoping to convey both the joys and realities of taking this approach to motivate those who haven't taken this path yet, whilst simultaneously building a shared connection with those who have.

Efficient Data Integration Techniques in Some Modern Applications*Kun Liu, JPMorgan Chase & Company*

Modern data science often requires integrating information from heterogeneous sources to support timely and reliable decision-making. This talk discusses efficient data integration techniques motivated by several modern applications, including online monitoring of large-scale data streams, consensus sequential detection over distributed networks, clinical trial analysis, and robust statistical inference with contaminated data. I will highlight how scalable statistical procedures can combine local information into useful global decisions, with emphasis on methods that remain computationally feasible, statistically efficient, and robust in complex real-world settings. The talk will also connect these methodological ideas to broader applications in finance, healthcare, manufacturing, and security, where data-driven decisions must often be made under uncertainty and at scale.

Slope-Aware RDP for Slope Changepoint Detection*Meenu Rani, Indian Institute of Technology Ropar*

This talk addresses the problem of detecting changes in the slope of a regularly observed time series. The Ramer-Douglas-Peucker method is commonly used to simplify curves by keeping the most important points that describe their overall shape. In this work, we introduce a slope-aware version of this method for identifying points where the trend of a time series changes. Instead of choosing a technical threshold, the proposed method uses only

the desired number of changepoints and then selects the most relevant points of slope change. The method is simple to use, computationally efficient, and can be applied directly to the original time series. In simulation studies, its ability to locate changepoints improves as the signal becomes clearer. One possible application is in financial data, where the method can help identify changes in the trend of stock prices, market indices, or other financial time series.

Online Change Point Detection Based on Machine Learning Approach*Ayten Yiğiter, Hacettepe University*

The change point problem arises in many fields, including financial time series analysis, signal processing, and statistical process control. It concerns the identification of structural changes in the statistical properties of sequentially observed data. Such changes typically arise from shifts in the parameters of the underlying data-generating process and may be reflected in changes in statistical moments or other characteristics of the observed series. Accurate detection of change points is of considerable importance, as it directly influences data modeling, forecasting accuracy, and subsequent decision-making processes.

In the literature, change point detection methods are generally categorized into two main settings: offline and online detection. In addition, different methodologies have been proposed depending on whether the underlying change occurs abruptly or evolves gradually over time. In this study, the online change point detection problem is investigated using an approach that incorporates machine learning components. The proposed approach is evaluated using real-world datasets, and its performance is compared with existing methods to demonstrate its effectiveness and practical applicability.



Session Info

S404

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 13:00 - 14:00 UTC

Coordinating Evidence for Better Research Design: Priors, Assumptions, and Scientific Judgment

USE SCHED TO ACCESS THE ZOOM ROOM

This session examines how scientific judgment can be made explicit in research design and decision making with mixed methods evidence, causal knowledge, expert elicitation, and Bayesian priors. Noor Qaragholi opens by showing how qualitative and quantitative evidence can inform strong, untestable causal assumptions in study design, illustrated through CACE estimation in a randomized behavioral intervention with substantial noncompliance. Utkarshani Jaimini then turns to Causal Neuro-Symbolic AI, a framework for building systems that are explainable, interpretable, and causally aware rather than limited to correlational pattern matching. Her work spans knowledge representation, causal inference, and applied machine learning. Anna Heath discusses expert elicitation for Bayesian clinical trial design, focusing on remote, real time exercises that help experts translate experience into probabilistic statements for prior distributions. Lu (Maggie) Qian closes by examining how Bayesian trial design can make scientific judgment explicit without allowing prior choice to become opaque, drawing on FDA guidance and reverse-Bayes methods as an audit tool. Together, the talks show how making assumptions and priors explicit can strengthen research design and support credible, decision relevant conclusions.

ORGANIZER: Elizabeth Eisenhauer, Westat**CHAIR:** Elizabeth Eisenhauer, Westat

Speaker Bios

S404



DR. NOOR QARAGHOLI
The Policy & Research Group

Noor Qaragholi is a public health researcher with 15 years of experience conducting qualitative, quantitative, and mixed methods research to improve health programs, services, and systems. As a Senior Research Analyst at The Policy & Research Group, Dr. Qaragholi leads large-scale evaluations and applied research studies for federal, state, and community partners, including projects focused on behavioral health, HIV, adolescent health, and infectious disease prevention. Her methodological work centers on integrating qualitative and mixed methods inquiry with causal inference, particularly to strengthen study design and generate the substantive knowledge needed to inform causal analyses that rely on strong, untestable assumptions. She earned her Doctor of Public Health in Implementation Science from the Johns Hopkins Bloomberg School of Public Health as a Bloomberg American Health Initiative Fellow and holds an MPH in Epidemiology and Biostatistics.

<https://utkarshani.github.io/>



DR. UTKARSHANI JAIMINI
University of Michigan - Dearborn

Utkarshani Jaimini is an Assistant Professor in the Department of Computer and Information Science at the University of Michigan-Dearborn, where she joined as a faculty in Fall 2025. She holds her doctoral degree from the University of South Carolina. Her research lies at the intersection of explainable AI and knowledge representation with a focus on leveraging causal knowledge graphs to make AI systems more transparent and interpretable. Her work has real-world impact across high-stakes domains including autonomous driving, smart manufacturing, and healthcare. She has a record of cross-sector engagement with contributions spanning academic publications and industry patents. Her research bridges the gap between foundational research and real-world application. She is a passionate advocate for interdisciplinary collaboration and continues to build partnerships that extend the reach and relevance of her work.

<https://evidenceinthewild.com/>



DR. LU QIAN
Zetyra, LLC

Lu (Maggie) Qian is an independent biostatistician and founder of Evidence in the Wild and Zetyra. Her work focuses on Bayesian and adaptive clinical trial design, regulatory alignment, prior specification, prior-data conflict, and operating characteristics of composed adaptive pipelines. She has submitted FDA comment letters on the 2026 draft Bayesian guidance and publishes the Evidence in the Wild essay series on Bayesian methods in clinical research; her writing has appeared in Significance. Her current collaboration with Robert A. J. Matthews develops fair-minded skeptical priors and reverse-Bayes regulatory diagnostics for Bayesian assessment of novel therapies. She will present an accepted contributed paper at JSM 2026 on validated, regulatory-aligned calculator design. She has more than 10 years of experience in oncology therapeutics and early detection clinical trials.

<https://lab.research.sickkids.ca/heath/>



DR. ANNA HEATH
The Hospital for Sick Children

Dr. Heath is a Scientist at The Hospital for Sick Children (SickKids), Toronto, an Associate Professor at University of Toronto and Honorary Research Fellow at University College London, UK. She is the Canada Research Chair in Statistical Trial Design. She arrived at SickKids in 2018 to lead the development of innovative statistical methodology for four clinical trials for novel treatments in the paediatric emergency department. In 2020, she founded the EMBaRC lab to focus on developing novel statistical methodology, software, expertise and guidance to improve the efficiency and design of randomised clinical trials.



Using Mixed Methods Evidence to Inform Scientific Judgment in Study Design

Noor Qaragholi, *The Policy & Research Group*

Establishing credible causal relationships is a central goal in science and medicine. Because estimating causal effects often depends on strong, untestable assumptions, researchers must exercise scientific judgment and draw on substantive knowledge to determine whether those assumptions are plausible in a given context. These judgements are integral for drawing credible causal conclusions. Yet, systematic approaches for generating and integrating the needed substantive knowledge into causal analyses remain limited.

One promising but underutilized strategy is the use of mixed methods – the purposeful integration of qualitative and quantitative approaches – to generate and incorporate the substantive knowledge relevant to causal assumptions. Such knowledge can strengthen study design by creating conditions under which assumptions are more likely to hold and can inform assessments of their plausibility. Using key assumptions underlying complier average causal effect (CACE) estimation as an illustrative example, Noor Qaragholi's presentation draws on a randomized trial of a behavioral intervention with substantial noncompliance. More broadly, this work discusses the potential for mixed methods approaches to strengthen study design, make scientific judgement more transparent, and strengthen confidence in causal conclusions.

Causal Neuro-Symbolic AI

Utkarshani Jaimini, *University of Michigan - Dearborn*

As AI systems become embedded in high-stakes decision-making, accuracy alone is no longer sufficient. We need systems that are explainable, interpretable, and causally aware. The current data-driven models are fundamentally correlational, identifying patterns without understanding the mechanisms behind them. This talk introduces Causal Neuro-Symbolic AI as a framework for bridging this gap combining causal reasoning with neuro-symbolic methods to help AI systems reason about why things happen, not just what tends to co-occur. The talk covers three threads: enriching neuro-symbolic AI with causal representations, integrating causal knowledge with domain expertise, and applying neuro-symbolic techniques to causal discovery. This work spans knowledge representation, causal inference, and applied machine learning.

How Bayesian Trial Design Can Make Scientific Judgment Explicit

Lu (Maggie) Qian, *Evidence in the Wild | Zetyra*

Bayesian clinical trial design inevitably requires scientific judgment, particularly in specifying prior distributions and determining how external evidence should influence inference. This talk examines how Bayesian methods can make those judgments explicit rather than opaque. Drawing on FDA Bayesian guidance, recent work on reverse-Bayes methods, and practical examples from clinical trial design, I discuss approaches for evaluating prior specification, detecting prior-data conflict, and

assessing dynamic borrowing in a transparent, reproducible manner. Rather than viewing priors as subjective inputs to be accepted on faith, these methods provide an auditable framework that strengthens scientific credibility while preserving the flexibility that makes Bayesian designs attractive.

Expert Elicitation for Bayesian Clinical Trial Design

Anna Heath, *The Hospital for Sick Children, University of Toronto, University College London*

Bayesian methods are increasingly popular in clinical research. However, the design of Bayesian clinical trials requires a prior distribution. One method to obtain these values is through expert elicitation, where experts are supported to turn their experience and expertise into probabilistic statements. Anna Heath and her team have used remote, real-time elicitation exercises to construct prior distributions to support the design of multiple clinical trials. This talk will introduce the challenges and advantages of using expert elicitation for clinical trial design.



Session Info



TECHNICAL STATISTICAL SESSION

CONTRIBUTED SESSION

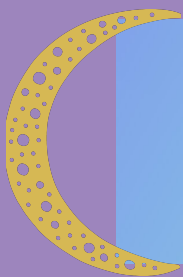
October 6th, 15:00 - 16:00 UTC

Selected Topics in AI and Data Science

USE SCHED TO ACCESS THE ZOOM ROOM

Big Data Is Not Neutral: Hidden Biases in Electronic Health Record Research | Spatio-Temporal Air Quality Risk Modelling | Interpretable Machine Learning-Based Variable Selection under Missing Data with the Feasible Solution Algorithm

S406

**ORGANIZER:** IDWSDS Program Committee**CHAIR:** Lúcia Henriques-Rodrigues, University of Lisbon

Speaker Bios

S406



DR. SIMA SHARGHI

Akron Children's Hospital

<https://www.linkedin.com/in/sima-sharghi/>

Dr. Sima Sharghi is a Senior Statistician and researcher with expertise in causal inference, Bayesian statistics, real-world evidence, electronic health record analytics, and predictive modeling in healthcare. She holds a PhD in Statistics and has extensive experience collaborating with multidisciplinary clinical and research teams on observational studies, longitudinal data analysis, simulation methods, and machine learning applications in medicine. Her work focuses on improving the rigor and interpretability of healthcare analytics through thoughtful study design, bias-aware statistical approaches, and the integration of Bayesian and causal inference methodologies in real-world clinical research. She has contributed to research involving pediatric health outcomes, social determinants of health, clinical prediction models, and methodological challenges associated with large healthcare datasets. She is passionate about advancing transparent and responsible data science in Medicine.



DR. MARÍA BUGALLO

University of A Coruña

<https://www.researchgate.net/profile/Maria-Bugallo>

María Bugallo is a Juan de la Cierva postdoctoral researcher at the University of A Coruña working in methodological statistics, with a focus on small area estimation and robust statistical modelling. She completed her PhD in Statistics at the Miguel Hernández University of Elche in 2024. Her research combines methodological developments with applications in official statistics, labour market analysis and environmental studies. She has published several articles in international journals and collaborates with the Spanish National Institute of Statistics on methodological projects. She has also presented her work at international conferences and completed research stays in Italy and Spain.



MS. MALIHA MEHNAZ MITU

University of Kentucky

Maliha Mehnaz Mitu is a PhD Candidate in Biostatistics at the University of Kentucky whose research focuses on methodological challenges in interpretable machine learning and Artificial Intelligence for healthcare. Her work integrates statistical learning, missing data methodology, and variable selection to improve the robustness, stability, and interpretability of clinical prediction models. Her current research examines how different missing data mechanisms and imputation strategies influence variable selection, selection stability, model performance, and reproducibility in electronic health record (EHR)-based machine learning.



Predictive Modeling for Irregular and Unbalanced Longitudinal Data: Application to Bariatric Surgery Outcomes

Kristina Vatcheva, University of Texas Rio Grande Valley

Metabolic and bariatric procedures, including Roux-en-Y gastric bypass and sleeve gastrectomy, produce substantial and durable weight loss, but postoperative trajectories vary widely. Predicting outcomes is challenging because longitudinal bariatric data are often unbalanced and irregularly timed. Traditional statistical models such as linear mixed-effects (LME), generalized estimating equations (GEE), and generalized additive mixed models are commonly used for correlated data but may struggle with complex nonlinear patterns. Machine learning approaches, including mixed-effects random forests, RE-EM trees, and neural networks (ANN, MLP, RNN, LSTM, GRU), offer greater flexibility.

In this study, neural networks best fit observed patient trajectories, effectively capturing nonlinear and temporal patterns. However, classical models, particularly LME and GEE, performed better on unseen test data, indicating stronger generalizability. These findings suggest that recurrent neural networks are well-suited for modeling patient history and forecasting future outcomes, while GEE provides more reliable predictions of current BMI for new patients.

Big Data Is Not Neutral: Hidden Biases in Electronic Health Record Research

Sima Sharghi, Akron Children's Hospital

Electronic health records (EHRs) are increasingly used for clinical research, predictive modeling, and real-world evidence generation. Although large healthcare datasets are often viewed as objective and comprehensive, they are shaped by healthcare utilization patterns, documentation practices, socioeconomic disparities, and measurement limitations. As a result, bias can persist — and sometimes amplify — despite very large sample sizes.

This presentation explores common hidden biases in EHR-based research, including missing data, outcome misclassification, differential healthcare access, temporal bias, and the use of geographic proxies for social determinants of health. Using simulation-based examples, the talk demonstrates how seemingly robust analyses may produce misleading conclusions when underlying data-generating mechanisms are ignored.

The session will also discuss the distinction between predictive performance and causal interpretation in healthcare analytics, emphasizing the importance of rigorous study design and statistical reasoning when working with observational clinical data. Rather than focusing on a single institution or proprietary dataset, this presentation highlights generalizable methodological challenges relevant across healthcare systems, industry, and academic research.

Spatio-Temporal Air Quality Risk Modelling

María Bugallo, University of A Coruña

This study introduces a robust spatio-temporal M-quantile framework for modelling air quality risk indicators in the presence of complex dependence structures and incomplete spatial coverage. By integrating Geographically Weighted Regression with dual spatio-temporal weighting schemes, the proposed methodology accommodates local distributional features and heterogeneous dynamics without relying on restrictive random-effects assumptions. Developed within the context of Small Area Estimation, the framework provides reliable out-of-sample prediction and accurate risk mapping for unmonitored domains, together with analytical estimators of the mean squared error. The performance of the proposed approach is assessed through extensive simulation studies and illustrated using U.S. county-level air quality data spanning the period 2016–2023. The empirical analysis identifies persistent pollution hotspots exceeding warning thresholds in several areas of California and southern Florida, highlighting the potential of the framework for environmental monitoring and public health risk assessment.

Interpretable Machine Learning-Based Variable Selection under Missing Data with the Feasible Solution Algorithm

Maliha Mehnaz Mitu, University of Kentucky

Modern machine learning models increasingly inform decisions that shape clinical and population health, yet these models are often built on incomplete data. In this study, we used simulation to evaluate the impact of different missing data handling strategies when applied with the Feasible Solution Algorithm (FSA), a machine learning method used for variable selection. For multiple imputation, we provide practical guidance using voting-based consensus rules to identify variables consistently selected across imputed datasets. Predictive performance was assessed using area under the curve, sensitivity, specificity, and accuracy on independent test sets. We also illustrate the approach with an application to real-world data using electronic health records to predict colorectal cancer screening.

Our results provide insights into when simple approaches, such as complete case analysis, may be adequate and when imputation-based strategies are necessary to improve stability and predictive performance. This work offers a practical framework for applying FSA in settings with missing data and informs broader use of automated variable selection methods in real-world health data.



Session Info

S407

TECHNICAL STATISTICAL SESSION

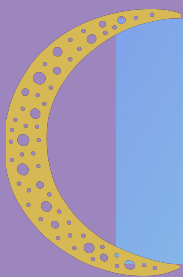
CONTRIBUTED SESSION

October 6th, 17:00 - 18:00 UTC

Selected Topics in Statistics 2

USE SCHED TO ACCESS THE ZOOM ROOM

Scalable Nearest-Neighbor Gaussian Graphical Models for Demographic Heterogeneity in
GLP-1 Outcomes|Paired Portfolio Trading: A Statistical Approach|Methodological
opportunities in genomic data analysis to advance health equity

ORGANIZER: IDWSDS Program Committee**CHAIR:** Jae Sook Cheong, Data Steward with Data Governance

Speaker Bios

S407



MS. GAOQIANXUE LIU
Drexel University

Gaoqianxue (Daphne) Liu is a PhD student in Biostatistics at the Dornsife School of Public Health, Drexel University. Her research includes scalable Bayesian methods for high-dimensional and graphical models in electronic health records, longitudinal and trajectory modeling, and machine learning for adolescent substance use and other health applications. She holds an MS in Biostatistics from Yale University and a BS in Mathematics from Shandong University.



MS. LEANDRA BRAEUNINGER
University College London

<https://leandrabraeuninger.github.io/>

Leandra Bräuninger (they/them; she/her) is a doctoral student at University College London, supervised by Dr Brieuc Lehmann and Prof. Ioanna Manolopoulou, using statistical and machine learning methods to define, quantify and mitigate genomic inequity. Drawing on algorithmic fairness, semi-parametric Statistics and Gaussian processes, conformal prediction, and applying it to topics such as equitable polygenic risk scores and variant effect prediction, Leandra's research leverages a variety of disciplines to solve complex problems at their intersection. Previously, Leandra held research and teaching positions in statistical fairness in genomics, mathematical malaria prediction, and fundamental biology at the Alan Turing Institute, the Mathematical Biology Group at the University of Melbourne, and the Centre for Active Learning at ETH Zurich. Leandra holds an MSc in Mathematical Modelling of Global Health (Oxford) and a BA in Data Science and International Relations (St. Gallen, CH).



MS. SARA MEZURI
Graduate Research Assistant

I am a Ph.D. candidate in the Department of Mathematics and Statistics at Oakland University in Michigan, specializing in Applied Mathematics. Currently, I am in my fifth and final year, and I expect to graduate in December 2026. My research involves developing several statistical frameworks and algorithms that combine canonical correlation analysis, homologous correlation, and cointegration testing to create and trade synthetic portfolio pairs, as well as predict using weekly returns. After graduation, I aspire to transition into the industry and pursue a role in finance or data science.





Scalable Nearest-Neighbor Gaussian Graphical Models for Demographic Heterogeneity in GLP-1 Outcomes

Gaoqianxue Liu, Drexel University

Treatment response to GLP-1 receptor agonists varies widely across demographic groups, yet the joint relationships among clinical biomarkers are rarely modeled across strata. Existing joint Gaussian graphical models couple groups through a global prior and become computationally costly as the number of strata grows. We introduce the Nearest-Neighbor Gaussian Graphical Model (NNGGM), which adapts the Vecchia approximation from spatial statistics, linking each demographic group mainly to its most similar neighbors. This substantially reduces computational cost and stabilizes estimation in small and unbalanced subgroups. Applied to a TriNetX cohort of chronic kidney disease patients initiating GLP-1 therapy, the method reveals substantial heterogeneity in biomarker networks and clear local structure across demographic strata. NNGGM offers a scalable and robust framework for studying how clinical dependency structure differs across populations. Although motivated by clinical biomarkers, the method applies broadly in settings where dependency networks are compared across many related groups, such as genomics, neuroimaging, and finance.

acknowledged that underrepresentation of diverse populations in human genetics research risks exacerbating existing health disparities. Efforts to improve diversity are ongoing, but an often-overlooked source of inequity is the choice of analytical methods used to process, analyse and interpret genomic data. This choice can influence all areas of genomic research, from genome-wide association studies and polygenic score development to variant prioritization and functional genomics. New statistical and machine learning techniques to understand, quantify and correct for the impact of biases in genomic data are emerging within the wider genomic research and genomic medicine ecosystems. At this crucial time point, it is important to clarify where improvements in methods and practices can, or cannot, have a role in improving equity in genomics. Here, we review existing approaches to promote equity and fairness in statistical analysis for genomics, and propose future methodological developments that are likely to yield the most impact for equity.

Paired Portfolio Trading: A Statistical Approach

Sara Mezuri, Graduate Research Assistant

Traditional pairs trading focuses on selecting a pair of individual stocks. However, in practice, it is difficult to identify pairs that may exhibit similar behavior over an extended period of time. In this work, we instead take a more drastic approach of paired portfolio trading, where, using the statistical techniques, instead of looking for pairs, we carefully construct the paired portfolios. We do so by using the multivariate techniques of canonical correlations and cointegration.

We apply and illustrate our approach on the daily data for ten large-cap U.S. equities from January 2021 to December 2023. We construct a host of paired portfolios from these ten equities, which have a high possibility of cointegrating and hence allow us the potential for profitable paired portfolio trading. Our strategy is almost market neutral in that the net investment of money in the market is minimal. We develop several entry and exit rules and refine our approach by using a variety of modelling techniques, such as orthogonal regression.

We also evaluate the strategy for the future out-of-sample data from January 2024 through December 2025. Our analysis shows that the strategy is profitable for all cointegrated pairs in-sample, achieving a 100% win rate, while also maintaining profitability out-of-sample.

Methodological opportunities in genomic data analysis to advance health equity

Leandra Braeuninger, University College London

The causes and consequences of inequities in genomic research and medicine are complex and widespread. However, it is widely



CAREER OR PROFESSIONAL DEVELOPMENT SESSION

PLENARY TALK

October 6th, 18:00 - 18:30 UTC

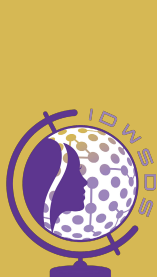
Bridging the Gap: How Practicing Data Scientists Use LLMs in the Wild and What It Means for Data Science Education

USE SCHED TO ACCESS THE ZOOM ROOM

Since the widespread availability of generative artificial intelligence (GenAI), particularly Large Language Models (LLMs), fundamental questions have emerged about the future of coding in data science. Some predict that data scientists will no longer need traditional coding skills, while others question whether LLMs might replace data scientists entirely. However, these discussions have largely proceeded without empirical evidence of how practicing data scientists actually use these tools.

This study addresses this gap by surveying trained, practicing data scientists to understand if and how they integrate LLMs into their workflows, particularly for writing and editing code and performing other data science tasks. Building on our recent investigation of data science educators' perspectives on LLMs, this research examines real-world usage patterns among practitioners to bridge the gap between current practice and educational preparation.

Our findings will contribute to the data science community in two critical ways. First, by documenting how data scientists are actually working with LLMs four years after their initial release, we provide actionable insights that allow practitioners to learn and adopt effective strategies for integrating these tools into their work. Second, we inform data science education by evaluating whether current pedagogies adequately prepare students for this evolving landscape.

ORGANIZER: Tiffany Timbers, University of British Columbia**CHAIR:** Yuyan Yi, NIAID

Speaker Bios

S408



DR. TIFFANY TIMBERS
University of British Columbia

<https://tiffanytimbers.com/>

Dr. Tiffany Timbers is an Associate Professor of Teaching in the Department of Statistics and Instructor in the Master of Data Science program at the University of British Columbia. She holds a PhD in Neuroscience from UBC and completed postdoctoral research in behavioral and neural genomics at Simon Fraser University, as well as a teaching Postdoctoral fellowship in Data Science at UBC. Dr. Timbers is co-author of three open educational resource textbooks, *Data Science: A First Introduction* (versions in R and Python) and *Python Packages*, all published through CRC Press. She is an internationally recognized leader in statistics and data science education, known for creating broadly accessible learning environments that welcome students from diverse backgrounds. In 2023, she received the Statistical Society of Canada's Early Career Educator Award for her innovative contributions to data science education.



Session Info

S409

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

PLENARY TALK

October 6th, 18:30 - 19:00 UTC

Every Step Counts: A Journey Through Crossroads, Turning Points and Learnings

USE SCHED TO ACCESS THE ZOOM ROOM

A career is rarely a straight line. It is shaped by choices, unexpected opportunities, challenges, mentors, setbacks, and the willingness to keep learning along the way. In this talk, I reflect on my journey from studying statistics in India to pursuing a PhD in Biostatistics in the United States and building a career as a statistician in the biopharmaceutical industry. Along the way, my experiences have taken me across academic research, internships, clinical development, statistical methodology, multiple therapeutic areas, mentoring, professional service, and leadership within the statistical community.

Rather than focusing only on milestones, this talk explores the crossroads and turning points behind them—the decisions that changed direction, the opportunities that initially seemed small but became important, and the lessons learned from navigating uncertainty and growth. Drawing from experiences in research, clinical trials, interdisciplinary collaboration, mentoring, and professional engagement, I will share how curiosity, adaptability, relationships, and continuous learning have shaped my development as a statistician. The central message is simple: careers are built one step at a time, and even the steps that do not seem significant in the moment can ultimately help define where we go next.

ORGANIZER: Arinjita Bhattacharyya, Inference Inc.**CHAIR:** Manushi Siriwardana, Penn State University

Speaker Bios

S409



DR. ARINJITA BHATTACHARYYA
Inference Inc.

Arinjita Bhattacharyya, PhD, is a biostatistician and drug developer with nearly a decade of experience across the pharmaceutical industry and academia. Most recently an Associate Principal Scientist in Biostatistics at Merck, she has supported clinical development across oncology, neuroscience, infectious diseases, cardiovascular disease, pediatrics, and other therapeutic areas. Her expertise includes clinical trial design, Bayesian and adaptive methods, survival analysis, statistical modeling, simulation, and regulatory submissions. She has extensive hands-on experience with R, SAS, R Shiny, and, including developing analytical tools and reproducible statistical workflows. Arinjita is also a published researcher, invited speaker, mentor, and active member of the statistical community, with multiple journal publications and extensive experience reviewing and editorial scientific research.



Session Info

S410

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 20:00 - 21:00 UTC

Statistical Inference with Incomplete Data

USE SCHED TO ACCESS THE ZOOM ROOM

Incomplete data arise in many scientific fields and present major challenges for modern statistical methodology. Missing values, censoring, and selection bias may substantially affect estimation, hypothesis testing, prediction, and the generalizability of scientific findings. Developing statistical methods that appropriately account for these sources of incompleteness is therefore essential for obtaining reliable conclusions. This session brings together recent methodological advances addressing different aspects of statistical inference with incomplete data. Topics include assessing the generalizability of randomized controlled trials through a sensitivity analysis framework for selection bias, evaluating modern imputation methods for recovering the underlying data distribution, independence testing for high-dimensional incomplete data, and nonparametric testing for censored survival data. Together, these contributions provide complementary perspectives on the analysis of incomplete data and demonstrate how principled statistical methodology can improve inference across a broad range of applications.

ORGANIZER: Marija Cuparić, University of Belgrade**CHAIR:** Marija Cuparić, University of Belgrade

Speaker Bios

S410



PROF. REBECCA ANDRIDGE
The Ohio State University

<https://cph.osu.edu/people/randridge>

Rebecca Andridge is a Professor of Biostatistics at The Ohio State University College of Public Health. She conducts methodologic work on imputation methods for missing data, primarily when missingness is driven by the missing values themselves (missing not at random), and on measures of selection bias for nonprobability samples. She teaches introductory graduate and undergraduate biostatistics, primarily to public health students, and won the Ohio State University's Alumni Award for Distinguished Teaching in 2026. She is an elected Fellow of the American Statistical Association.



MS. KRYSZYNA GRZESIAK
University of Wrocław

<https://github.com/KrystynaGrzesiak>

Krystyna Grzesiak is a PhD candidate at the Institute of Mathematics, University of Wrocław. She conducts research on developing machine learning methods for the analysis of proteomic and metabolomic data.



PROF. BOJANA MILOŠEVIĆ
University of Belgrade

<https://poincare.matf.bg.ac.rs/~bojana/en/>

Bojana Milošević is an Associate Professor and Head of the Department of Probability and Statistics at the Faculty of Mathematics, University of Belgrade. Her research interests include nonparametric statistics, statistical hypothesis testing, survival analysis, dependence analysis, reliability theory, and statistical inference with incomplete data. She received her PhD in 2016. She has published 48 scientific papers and presented her research at more than 85 international conferences, including 29 invited talks. She is the author of two university textbooks and has supervised 45 master's theses and 3 PhD dissertations. She has led several national and international research projects. She is an elected member of the International Statistical Institute (ISI), Editor of *Bernoulli News*, serves on the editorial boards of several scientific journals, and is a member of the Management Committee of COST Action CA21163 HiTEC.



MS. ANKE STEYN
North-West University

<https://natural-sciences.nwu.ac.za/centre-business-mathematics-and-informatics/anke-steyn>

Anke Steyn is a lecturer at the North-West University in Potchefstroom, South Africa. Her research interests include survival analysis, goodness-of-fit testing and experimental design. She qualified as an actuary and recently completed her doctoral degree in Statistics. Her great passions in life are to make sense of data, and to teach so that the 'lightbulb' clicks on. In her free time, she enjoys baking and spending time in nature.



Selection Bias in Experimental Results: A Proxy Pattern-Mixture Model for Randomized Trial Generalizability

Rebecca Andridge, The Ohio State University College of Public Health

Randomized controlled trials (RCTs) are the gold standard for assessing interventions, yet most use convenience samples rather than random samples from a target population. If unobserved differences between participants and non-participants include factors related to treatment effectiveness ("effect modifiers"), trial results may fail to generalize, contributing to the "efficacy-effectiveness gap". In this talk, I introduce a novel sensitivity analysis framework to assess the impact of such unmeasured factors on treatment effect estimates called the Proxy Pattern-Mixture Model in the context of RCTs (RCT-PPMM). This framework that treats this selection bias as a fundamental problem of incomplete or missing data because outcomes for the entire target population are not observed. The RCT-PPMM relies on two bounded sensitivity parameters that capture the deviation from sample selection at random and that can be varied systematically to determine how robust trial results are to a departure from ignorable sample selection. A critical advantage is that it requires only summary-level baseline covariate data for the target population, rather than individual-level data for nonparticipants. I illustrate the use of the method using a yoga intervention RCT for breast cancer survivors, showing how conclusions may shift under plausible selection biases. The approach offers a practical and interpretable tool for evaluating generalizability, particularly when individual-level data on nonpart

matters. We focus on two related problems: testing mutual independence among several variables and testing independence between two random vectors in the presence of missing observations. Particular attention is given to the impact of different missingness mechanisms, such as MCAR and MAR, on test statistics, asymptotic behaviour, and finite-sample performance. We compare several approaches, including complete-case procedures, weighting-based methods, and imputation strategies, and provide practical recommendations for different missing-data scenarios.

On the k-sample problem in the presence of random right censoring

Anke Steyn, Centre for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa

Across different disciplines and fields, a question that often arise is whether two or more samples of time-to-event data are realised from the same distribution. The problem is complicated by the presence of random right censoring. We discuss a new test based on empirical characteristic functions. The proposed test statistic is a weighted L2 type distance between the empirical characteristic functions of the observed samples. In the presence of censoring, the empirical characteristic function is based on the Kaplan-Meier estimator. We investigate the asymptotic properties of the test, and consider the finite-sample performance of the test for different censoring proportions, and $k=2,3$ and 5.

A comprehensive evaluation of missing data imputation methods

Krystyna Grzesiak, University of Wrocław, Faculty of Mathematics and Computer Science

Missing values are ubiquitous in many real-world datasets. While imputation methods are typically evaluated through pointwise errors, the primary goal of imputation is to recover the underlying data distribution. In this work, we present a comprehensive benchmark of modern imputation methods, comparing their ability to reconstruct the original data distribution across a wide range of datasets and missingness mechanisms.

Independence Testing with Incomplete Data: When Missingness Matters

Bojana Milošević, Faculty of Mathematics, University of Belgrade

High-dimensional data are increasingly common in modern applications, including genomics, biomedical studies, and environmental monitoring. In such settings, testing independence is already challenging due to the large number of variables, complex dependence structures, and limited sample sizes. These difficulties become even more pronounced when the data are incomplete. Missingness may substantially affect both the validity and the power of independence tests, while naive imputation may distort the underlying dependence structure.

In this talk, we discuss independence testing for high-dimensional incomplete data and examine when and how missingness



Session Info

S411

OTHER

INVITED PANEL

October 6th, 21:00 - 22:00 UTC

Fireside Chat with CWS Country Representatives: Navigating the Impact of AI in Teaching, Research, and Professional Practice

USE SCHED TO ACCESS THE ZOOM ROOM

Artificial Intelligence (AI) is transforming higher education and professional practice at an unprecedented pace, reshaping teaching, research, and ethics while shifting the professional landscape for women and underrepresented groups in data science. This invited fireside chat brings together CWS country representatives to share insights from diverse global contexts, highlighting how regional infrastructure gaps and regulatory frameworks shape AI adoption. Structured as an interactive, ask-me-anything-style conversation, the session will engage the audience to foster open dialogue. Through dynamic exchanges, panellists will explore critical tensions in teaching and learning - such as adapting course design and student assessments to an AI-enabled landscape while combatting the threat of professional "deskilling". The discussion will also examine the dual-edged sword of AI in research, balancing efficiency gains in coding and data processing against concerns regarding algorithmic bias, automated plagiarism, and transparency. Moving beyond academia, the session will address how AI influences broader professional practice, focusing on the need for inclusive ethical frameworks and policies that promote equity. Ultimately, this cross-regional dialogue aims to identify practical strategies for responsible AI use, fostering global awareness of the opportunities and systemic challenges faced by educators, researchers, and professionals in statistics and data science.

ORGANIZER: Vanda Lourenço, NOVA School of Science and Technology

CHAIR: Wendy Lou, Dalla Lana School of Public Health, University of Toronto

SPONSOR: MRC Biostatistics Unit, University of Cambridge



Speaker Bios

S411



DR. LÍGIA HENRIQUES-RODRIGUES
University of Lisbon

Lígia Henriques Rodrigues is an Associate Professor at the Faculty of Sciences of the University of Lisbon. She is a researcher at the Centre of Statistics and Its Applications of the University of Lisbon.

Her research focuses mainly on Extreme Value Statistics, semi-parametric estimation, bias reduction, risk modelling, Bayesian analysis, spatio-temporal models and survival analysis. Her work combines methodological developments in Statistics with applications to real data, particularly in the analysis of extremes, risk and lifetime data.

She has been involved in several research projects and scientific events. She is also a board member of the Portuguese Statistical Society and of PT-MATHS-IN, and currently serves as President of the European Courses in Advanced Statistics (ECAS - FENStatS).



DR. GABRIELA CYBIS
Federal University of Rio Grande do Sul, Brazil

Gabriela Bettella Cybis is an Associate Professor in the Department of Statistics at the Federal University of Rio Grande do Sul (UFRGS), Brazil, where she is also a faculty member of the Graduate Program in Statistics.

She holds a B.Sc. in Biology and an M.Sc. in Mathematics from UFRGS, and a Ph.D. in Biomathematics from the University of California, Los Angeles (UCLA).

Her research focuses on Bayesian inference and computational methods for genetics, evolution, phylodynamics, and viral epidemiology, as well as nonparametric methods for classification and clustering. She is currently supported by a joint Serrapilheira Institute and FAPERGS research grant on machine learning for phylodynamics and statistical methods for high-dimensional data.



DR. ATINUKE ADEBAJJI
Purdue University

Atinuke Adebajji is an Associate Professor of Practice- Statistics at Purdue University, Indianapolis.

Her research and engagement activities focuses on applied multivariate methods, statistical modelling of public health outcomes, multidisciplinary statistical collaboration, and promoting statistical literacy. She has published articles on new distributions applied to health data, classification and pattern recognition, and transportation network. Her research has resulted in more than 90 scientific papers in collaboration with co-authors from different countries and delivered more than 100 invited talks at public forums, conferences and scientific seminars.

Atinuke is actively engaged in career and life coaching. She is a Fellow of the Ghana Academy of Arts and Sciences, an Elected member of the International Statistical Institute and, a Member of the Representative Council of the International Biometric Society.



DR. SOFÍA VILLAR
University of Cambridge

Sofia S. Villar is a Research Professor and Group Leader within the Efficient Study Design theme at the MRC Biostatistics Unit, University of Cambridge, where she also serves as Academic Lead for Equality, Diversity and Inclusion. Her research bridges optimization, machine learning, and statistics to develop computationally feasible, patient-centric adaptive trial designs that accelerate treatment delivery while preserving statistical power. She holds a PhD in Business Administration and Quantitative Methods (Universidad Carlos III de Madrid, 2012) and was awarded the prestigious Biometrika post-doctoral fellowship in 2014. Dr. Villar co-leads the Adaptive Designs Working Group for the MRC-NIHR Trials Methodology Research Partnership. As senior statistician, she steers major national innovative trials—including PIPAH, NOTACS, and STRATOSPHERE—implementing advanced Bayesian methods, sample size re-estimation, and response-adaptive randomization."



Session Info

S412

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

PLENARY TALK

October 6th, 22:00 - 22:30 UTC

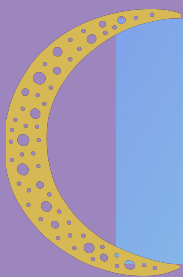
Building Trustworthy Financial Systems with Explainable AI: Applications in Credit Risk Modeling and Financial Inclusion

USE SCHED TO ACCESS THE ZOOM ROOM

Artificial intelligence is transforming how financial institutions assess creditworthiness, detect fraud, and expand access to financial services. Yet many of today's most powerful machine learning models remain opaque, making consequential decisions that customers, regulators, and even financial institutions struggle to interpret. As AI adoption accelerates, ensuring transparency, accountability, and fairness has become essential to building public trust and achieving equitable financial outcomes.

ORGANIZER: Angela Omogbeme, University of West Georgia

CHAIR: Denise Silva, Science- Sociedade Para O Desenvolvimento Da Pesquisa Cientifica



Speaker Bios

S412



MS. ANGELA OMOGBEME
University of West Georgia

<https://www.linkedin.com/in/angelaomogbeme/>

Angela Omogbeme is a financial technology researcher and data analytics professional specializing in artificial intelligence, explainable AI, fraud detection, credit risk modeling, and financial inclusion. She holds an M.S. in Business Analytics (4.0 GPA) from the University of West Georgia and an MBA from Edinburgh Business School, Heriot-Watt University, UK. Her research has been presented internationally, including at conferences hosted at the University of Oxford, UK and the University of London, where she received Best Presentation awards. Her peer-reviewed publications have been cited more than 130 times. Angela is a SCORE Certified Mentor and Industry Subject Matter Expert with the U.S. Small Business Administration, a Senior Member of IEEE who also serves as the UK Ambassador for GTA Black Women in Tech, a global organization partnered with UN Women. Her current research focuses on trustworthy and explainable AI for high-stakes decision-making in regulated industry.



Session Info

S500

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 2:00 - 3:00 UTC

Modern Frontiers in Complex Data Analysis: Scalable Bayesian Methods and Dynamic Online Estimation

USE SCHED TO ACCESS THE ZOOM ROOM

This session brings together pioneering research that tackles the dual challenges of complexity and scalability in modern statistics and data science. The presentations feature cutting-edge developments in Bayesian hypothesis testing, discrete choice modeling, and streaming spatial data analysis over complex domains. The session begins with two complementary talks introducing highly efficient, closed-form Bayesian alternatives to classical multivariate and random effects testing frameworks. It then transitions to an innovative Bayesian best-worst choice model that utilizes Integrated Nested Laplace Approximation (INLA) for complex preference data analysis. Concluding the discussion is a state-of-the-art online nonparametric estimation of spatially varying coefficient models for streaming spatial data using dynamic bivariate penalized splines over triangulations. Together, these talks highlight the vital role of advanced computational methodologies, spanning objective Bayes, deterministic approximations, and sequential updating algorithms, in achieving rigorous inference and extracting robust insights from complex data structures across diverse scientific domains.

ORGANIZER: Mai Dao, Wichita State University**CHAIR:** Mai Dao, Wichita State University

Speaker Bios

S500



DR. ZHUANZHUAN MA
The University of Texas Rio Grande Valley

https://scholar.google.com/citations?hl=en&user=PhGRp6UAAAAJ&view_op=list_works&sortby=pubdate

Dr. Zhuanzhuan Ma is an Assistant Professor of Statistics in the School of Mathematical and Statistical Sciences at the University of Texas Rio Grande Valley. She earned her Ph.D. in Statistics from Texas Tech University in 2022. Her research interests include Bayesian statistics, high-dimensional inference, machine learning, and their applications in biomedical and health-related studies. Her current work focuses on developing innovative statistical methodologies for complex and large-scale data. Outside of academia, she enjoys CrossFit and badminton.



DR. MAI DAO
Wichita State University

https://www.wichita.edu/profiles/academics/fairmount_college_of_liberal_arts_and_sciences/Math/Faculty/Dao-Mai.php

Dr. Mai Dao is an Assistant Professor of Statistics in the Department of Mathematics, Statistics, and Physics at Wichita State University. She obtained her Ph.D. from Texas Tech University under the supervision of Professors Min Wang (The University of Texas at San Antonio) and Souparno Ghosh (University of Nebraska-Lincoln). Her primary research interests focus on Bayesian statistics, high-dimensional inference, and statistical machine learning.



DR. NADEESHA JAYAWEERA
University of Akron

<https://www.uakron.edu/statistics/faculty-staff/bio-detail.dot?u=njayaweera>

Dr. Nadeesha Jayaweera is an Assistant Professor in the Department of Statistics at the University of Akron. She received her Ph.D. in Statistics from Texas Tech University in 2022 and previously held a postdoctoral research position at Worcester Polytechnic Institute. Her research interests include Bayesian statistics, spatio-temporal modeling, discrete choice analysis, time series analysis, sequential analysis, statistical machine learning, and disease surveillance.



DR. JINGRU MU
Kansas State University

<https://www.k-state.edu/stats/about/people/MuJingru.html>

Dr. Jingru Mu is an Associate Professor in the Department of Statistics at Kansas State University. Her research focuses on developing flexible and computationally efficient statistical methods for complex, large-scale, and spatio-temporal data. Her interests include nonparametric and semiparametric inference, variable selection, and image analysis. She earned her Ph.D. in Statistics from Iowa State University and is dedicated to research, teaching, and mentoring in statistics and data science.



A Simple Two-Sample Bayesian Hotelling's T^2 for Equality of Means*Zhuanzhuan Ma, The University of Texas Rio Grande Valley*

We propose an explicit Bayes factor for comparing the mean vectors of two populations in multivariate hypothesis testing. The proposed approach could be viewed as a natural Bayesian counterpart to the classical Hotelling's T^2 test, thereby connecting Bayesian and frequentist perspectives. A notable advantage of the present procedure is that it depends on the observed data only through the Hotelling's T^2 statistic, which leads to computational simplicity and intuitive interpretation, and makes it particularly suitable for inclusion in multivariate analysis courses that emphasize Bayesian thinking. We then investigate the desiderata of the proposed Bayesian procedure: information consistency and predictive matching, and also study its model selection consistency under two different asymptotic scenarios: when the number of units tends to infinity and when the number of observations per unit tends to infinity. Simulation studies and a real-data application are provided to illustrate the finite-sample performance of the proposed method.

A Consecutive Bayesian Hypothesis Test for the One-Way Random Effects Model*Mai Dao, Wichita State University*

In this talk, we discuss an objective Bayesian hypothesis testing procedure within the framework of one-way random effects models. Objective Bayesian methods typically rely on objective improper priors, which can lead to an undefined Bayes factor due to arbitrary constants. A common solution averages Bayes factors over all possible minimum training samples (MTS) to eliminate this constant, but incurs high computational costs when MTS combinations are large. To overcome this, we introduce a consecutive MTS size strategy to derive the consecutive Bayes factor (CBF), providing a simple, closed-form, and computationally efficient expression. The proposed CBF facilitates the computation of consecutive posterior probabilities, enabling an objective quantification of evidence against the null hypothesis. Furthermore, we examine the model selection properties of the CBF under two asymptotic scenarios: (i) as the number of units tends to infinity, and (ii) as the number of observations per unit tends to infinity. Simulation studies and real-data applications demonstrate the effectiveness of the proposed method.

Bayesian Modeling of Dependent Best-Worst Choices Using INLA*Nadeesha Jayaweera, The University of Akron*

We propose a Bayesian framework for analyzing highly dependent best-worst (BW) choice data in discrete choice experiments. To capture dependence between best and worst selections, BW outcomes are modeled as directed transitions between latent preference states. Transition counts are analyzed using a Poisson log-linear model with pair-specific random effects to account for heterogeneity and cross-alternative dependence. Bayesian inference is conducted using Integrated Nested Laplace

Approximation (INLA), providing fast and scalable estimation without requiring Markov chain Monte Carlo (MCMC). The proposed approach is evaluated through simulation studies and a quality-of-life case study based on Flynn et al. and compared with existing copula-based transition models. Results demonstrate improved model fit, stability across different sample sizes, and reliable estimation of latent utilities based on DIC and WAIC criteria. Overall, this framework provides a flexible, interpretable, and computationally efficient approach for analyzing complex preference data with applications in health, marketing, and social sciences.

Online Estimation of Spatially Varying Coefficient Models for Streaming Data over Complex Domains*Jingru Mu, Kansas State University*

Spatial data analysis has attracted considerable interest; however, it faces new challenges due to the increasing volume of spatial data streams. In this paper, we develop a novel online nonparametric method to update the estimates of coefficient functions in spatially varying coefficient models (SVCs). The updated estimation process is based on dynamic bivariate penalized splines over triangulations. The proposed approach features a gradient-magnitude-driven vertex insertion strategy that dynamically refines the triangulation over the complex domain following the Delaunay criterion, thereby periodically introducing more Bernstein basis functions to enhance the approximation of the online estimators as new data continuously flow in. Meanwhile, we design an efficient sequential updating algorithm that relies only on summary statistics and newly observed data to implement the proposed online estimation. Under mild regularity conditions, we establish the optimal L_2 convergence rate of the proposed online estimators. Furthermore, we construct multiplier-bootstrap-based online pointwise confidence intervals for the coefficient functions in the SVCs. Simulation studies and an empirical example demonstrate that the proposed online method performs nearly as efficiently as the classical methods that have access to the historical dataset.



Session Info

S501

TECHNICAL STATISTICAL SESSION

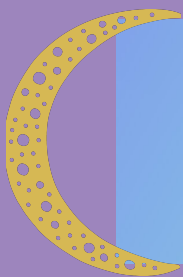
CONTRIBUTED SESSION

October 6th, 12:00 - 13:00 UTC

Additional Topics in AI and Data Science 1

USE SCHED TO ACCESS THE ZOOM ROOM

Graph-Aware Sparse Regression for Block-Missing Multimodal Neuroimaging Data |
Decoding Antarctic Microbial Biodiversity through Hybrid Machine Learning and
Constrained Ordination | Assessment of Methodological Quality in Non-Commercial
Clinical Studies: The Impact of International Reporting Guidelines on Studies
Approved | Faculty Teaching and Research Productivity and Equity in STEM: Student
Grade Gaps in Lower-Division Courses

ORGANIZER: IDWSDS Program Committee**CHAIR:** Victoria G. Tanoh, North Carolina A&T State University

Speaker Bios

S501



DR. ANNA KYE
University of California Irvine

Anna Kye is a Postdoctoral Scholar at UC Irvine, with appointments at the Postsecondary Education Research Institute and the Division of Teaching Excellence and Innovation. She holds a Ph.D. in Research Methodology (Quantitative) from Loyola University Chicago and an M.S. in Statistics from UMass Amherst. Her research applies cluster analysis, machine learning, and regression modeling to equity and faculty productivity in higher education. Her dissertation, awarded with distinction, compared four machine learning algorithms for college enrollment prediction using data from over 108,000 students. Prior to her postdoc, she spent seven years as an enrollment research analyst at Loyola University Chicago. One of her studies has been featured in FutureEd's Top 10 Research Notes of 2025. She has presented her research at JSM, WSDS, and AERA. She is a member of the American Statistical Association and the Caucus for Women in Statistics.

<https://it.linkedin.com/in/gabriella-tarantini-329a44130>



DR. GABRIELLA TARANTINI
IRCCS Azienda Ospedaliero Universitaria di Bologna

RESEARCH FELLOWSHIP & THESIS PROJECT, McMaster University, Department of Health Research Methods, Evidence, and Impact (HEI), Hamilton, Canada, from JUN/2026 to OCT/2026. Developed and implemented statistical models to identify and evaluate predictive factors of methodological quality in non-commercial clinical trials; Assessed study coherence and reporting compliance against international guidelines (such as CONSORT, TREND, and STROBE). HUB STUDY MANAGEMENT AND COORDINATION, CLINICAL TRIAL CENTER, IRCCS AOUBO – Bologna, Italy. Senior Study Coordinator in Clinical Trial Center Unit: Coordination and Management of Clinical Studies sponsored by IRCCS AOUBO, Support the design of clinical research and data analysis; Handling and management of clinical data for reporting to Italian Ministry of Health; Member of Clinical Trial Quality team. Recidency in Health Statistics and Biometrics, Ongoing, University of Pisa, Pisa, Italy.

<https://orcid.org/0009-0006-7173-8231>



MRS. MUTIYAT USMAN
North Carolina A&T State University

Mutiyat Adeola Usman is a Doctoral Researcher in Data Science and Analytics at North Carolina A&T State University, specializing in quantitative methodology, statistical modeling, and machine learning for complex, high-dimensional data. Her research focuses on developing Dynamic Gaussian Graphical Models to study evolving connectivity patterns in brain organoids and neural systems. She has authored peer-reviewed publications in statistical methodology, neuroengineering, public health, and artificial intelligence, with a recent submission to NeurIPS. Her expertise spans graphical models, longitudinal analysis, survey-weighted machine learning, network science, and reproducible data science. Passionate about translating methodological innovation into real-world impact, she develops scalable analytical frameworks that advance healthcare, neuroscience, and public health through rigorous, interpretable, and responsible AI.



MISS SILVIA MARÍA HERRANZ HERNÁNDEZ
Universidad Autónoma de Madrid

I graduated with two bachelor's degrees, in Mathematics and Computer Science, from Universidad Autónoma de Madrid, having also studied at the Università degli Studi di Torino. In September 2026 I started an MSc in Artificial Intelligence for Biomedicine and Healthcare at University College London. My research experience spans neuroscience, oncology and biomedical AI. I completed research internships at the Department of Systems and Functional Neuroscience, CSIC (Madrid) and at the Molecular Imaging and Radiation Oncology Lab at UCLouvain (Brussels), where I contributed to projects in neural data analysis and the standardisation of AI model reporting in radiotherapy. I co-authored a publication in phiRO journal. I currently work as a Technical Trainee at Siemens. My research interests lie at the intersection of statistics, machine learning, and the life sciences, with a particular focus on biomedical AI, AI-driven drug discovery, and trustworthy AI for healthcare applications.



Graph-Aware Sparse Regression for Block-Missing Multimodal Neuroimaging Data

Mutiyat Usman Mutiyat Usman, North Carolina A&T State University

Multimodal neuroimaging studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), frequently exhibit blockwise missing data, where subjects contribute only a subset of imaging and biomarker modalities. This structured missingness limits conventional graph-aware regression methods, which assume complete observations and a well-defined covariance matrix. Existing moment-based estimators, such as DISCOM, exploit all available paired observations without imputation but often yield covariance matrices that are not positive definite because covariance entries are estimated from heterogeneous sample sizes, preventing stable inverse-based graph regularization. We propose **Graph-Aware Sparse Regression for Block-Missing Multimodal Neuroimaging Data (GASR-BM)**, a unified framework that operates entirely at the covariance level. The method constructs a covariance estimator from all available paired observations, applies a positive definite correction to ensure stability and invertibility, and integrates graph-guided regularization to exploit anatomical and functional relationships among brain regions. By encouraging smooth coefficient estimation across connected features, GASR-BM identifies biologically meaningful subnetworks while accommodating incomplete multimodal observations. The proposed framework unifies covariance estimation, positive definite regularization, and graph-aware sparse regression into a scalable methodology for high-dimensional neuroimaging analysis.

Decoding Antarctic Microbial Biodiversity through Hybrid Machine Learning and Constrained Ordination

Silvia Maria Herranz Hernández, Universidad Autónoma de Madrid

In microbial ecology, the study of community diversity often involves ordering samples from biological communities in a space of latent variables that integrate information from observed environmental variables. In this context, Redundancy Analysis (RDA) is one of the most widely used multivariate technique (Legendre and Legendre, 2012). RDA extends Principal Component Analysis into a constrained ordination framework based on multivariate linear regression. However, it should not be applied to species abundance matrices, particularly when many zeros are present, since the assumptions of linear regression are violated, and it imposes the restrictive condition that the relationship between communities and environmental conditions be linear. We propose a novel hybrid framework, termed **hyb-RDA**, that integrates constrained ordination and machine learning to address some of the limitations of traditional RDA when analysing complex ecological datasets. Using **hyb-RDA**, we identify patterns in the composition of airborne microbial communities in Antarctica arising from local meteorological factors and the back trajectories of air masses. By capturing nonlinear relationships while preserving interpretability, **hyb-RDA** highlights the potential of hybrid ML-ordination approaches for analysing microbial biodiversity under complex environmental conditions.

Assessment of Methodological Quality in Non-Commercial Clinical Studies: The Impact of International Reporting Guidelines on Studies Approved

Gabriella Tarantini, IRCCS Azienda Ospedaliero Universitaria di Bologna

Non-commercial clinical trials are pivotal for advancing evidence-based medicine, yet their impact heavily relies on methodological robustness, transparency, and reproducibility. While Institutional Review Boards and Ethics Committees (IRBs/ECs) serve as primary gatekeepers of study quality, systematic assessments of protocol efficiency and reporting fidelity remain scarce. This study evaluates the efficiency and quality of non-commercial clinical protocols, assessing their administrative trajectory, adherence to protocol-specified publication timelines, and eventual reporting fidelity against international standard guidelines. I analyzed a comprehensive sample of nearly 200 clinical protocols spanning both interventional and observational studies. To enhance generalizability, the sample was drawn from two distinct international regulatory frameworks and institutional settings: IRCCS Azienda Ospedaliero Universitaria di Bologna in Italy and the McMaster University Research Ethics Board (HIREB) in Canada. I assessed the adherence between the statements in the original protocol and the final published article, specifically focusing on discrepancies in primary objectives, pre-specified endpoints, and sample size calculations. Moreover I evaluated the completeness of the final publications using international reporting frameworks tailored to the study designs, specifically CONSORT for RCT, TREND for non-randomized evaluations, and STROBE for observational studies.

Faculty Teaching and Research Productivity and Equity in STEM: Student Grade Gaps in Lower-Division Courses

Anna Kye, University of California Irvine

This study extends prior work that derived three STEM faculty academic productivity clusters from 16 standardized metrics using NbClust in R and validated them with ANOVA and Kruskal-Wallis tests. Those clusters are used here to examine equity in course outcomes among 553 faculty at a public R1 university. Average grade gaps between minoritized and non-minoritized students served as equity outcomes across four student groups. Four regression models were estimated with cluster membership, faculty characteristics, and course features as predictors. Tukey HSD tests confirmed significant gap variation across clusters. Faculty with moderate research productivity and heavy lower-division teaching loads were associated with smaller gaps for first-generation, low-income, and female students. Large class sizes widened gaps. The study demonstrates a pipeline combining cluster profiles, hypothesis testing, and regression to surface how faculty productivity shapes equity in STEM course outcomes.



Session Info

S502

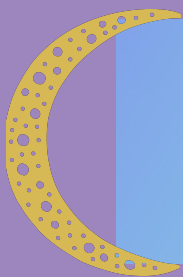
CONTRIBUTED SESSION

October 6th, 13:00 - 14:00 UTC

Additional Topics in AI and Data Science 2

USE SCHED TO ACCESS THE ZOOM ROOM

Developing and Using Data Visualization Requirements Gathering Processes for Clients | Mapping Global Open-Source Collaboration: A Data Science Approach to Measuring International Software Development | Multimodal AI Fusion for Better Alzheimer Detection | End-of-Life Healthcare Utilization among Medicare Beneficiaries with Parkinson's Disease Dementia at USA Academic Movement Disorders Centers: A 2019 Pre-Intervention Assessment

ORGANIZER: IDWSDS Program Committee**CHAIR:** Camilla Salvatore, Utrecht University

Speaker Bios

S502



DR. CHIPO ZIDANA
ZIMYAS

Dr. Chipó Zidana is a senior statistician, bio-statistical consultant, and data integration specialist with over 15 years of experience bridging advanced analytical modeling with national policy frameworks. She holds a PhD in Statistics from Çukurova University, specializing in mixture models for heterogeneous populations. Dr. Zidana's expertise spans multimodal data fusion, biomedical data engineering, geospatial predictive modeling, and national data harmonization alongside international partners like UNICEF.

A BQA-accredited lead moderator and assessor, she has served as Coordinator of the Statistical Services Unit at BIUST and Program Coordinator at NUST. Dr. Zidana is a prominent leader in regional capacity building and gender equity in STEM, serving as the Chairwoman of African Women in Data Science, a Fellow and Co-Chair of the Zimbabwe Young Academy of Sciences, and the Scientific Chair for the IBS SUSAN 2027 conference.

<https://sph.uth.edu/faculty/?fac=86IAK4QbEovkCDrymIVg5g==>



DR. ADRIANA PEREZ
The University of Texas Health Science Center at Houston

Professor at the Department of Biostatistics and Data Science. Dr. Pérez is engaged in a wide range of research projects: theoretical model evaluation accounting for imputation uncertainty; fitting complex data; analysis of cluster-randomized community trials; clinical trials; analysis of food intake involving measurement error; tobacco regulatory science; and implementation science. Dr. Pérez has expertise in conducting and disseminating research findings from youth and adults, as well as developing new statistical methodologies. She has completed and disseminated the results of several NIH grants. Dr. Pérez has over 150 peer-reviewed publications. Dr. Pérez is an active member of the International Biometric Society (IBS) and the American Statistical Association (ASA). She was the elected president of the Central American and Caribbean Region (RCAC) of IBS. Currently, Dr. Pérez serves as the representative council for RCAC IBS and chair of Inter-Region/Network Funding Task Force.



MISS SARAH GLIDDEN
Westat

Sarah Glidden is a research and data visualization analyst. She specializes in dashboard development for a range of external clients using Tableau.



MISS RASHI SALUJA
Westat

Rashi is a Data Scientist at Westat, where she works across a broad portfolio of research projects spanning survey methodology, public health, applied machine learning, and artificial intelligence. Her recent work includes network analysis of open-source software collaboration, as well as developing LLM evaluation frameworks and adaptive survey design methods for federal research programs. Rashi is passionate about translating complex technical solutions into clear, actionable insights for clients and collaborators. Outside of work, she enjoys hiking, cooking, and following emerging developments in AI.



Developing and Using Data Visualization Requirements Gathering Processes for Clients

Sarah Glidden, Westat

Creating data visualizations and dashboards for clients' decision-making needs requires an iterative requirement gathering process before actual visualization development can begin. Developing consistent requirement gathering steps that are used across all data visualization projects ensures consistency and accuracy as a data visualization team. An interview with the client is the first step to assessing their needs and the properties of the data that will be used, as well as understanding their level of experience communicating about visualizations. Following this interview, formal write-ups and static draft visualizations can be iterated on between the data visualization analyst and the client. This talk will share processes, strategies, and challenges from data visualization requirements gathering that may benefit data visualization analysts working with a diverse array of clients.

Mapping Global Open-Source Collaboration: A Data Science Approach to Measuring International Software Development

Rashi Saluja, Westat

Open-source software (OSS) has emerged as a critical indicator of a nation's science and technology capacity. Understanding who contributes to OSS offers valuable insight into the structure of global innovation. This study presents a data-driven framework for measuring international collaboration in open-source software development, developed as part of the 2026 Science and Engineering Indicators report, Translation to Impact: U.S. and Global Science, Technology, and Innovation Output, published by the National Science Board, at the U.S. National Science Foundation. OSS repositories serve as rich, observable records of developer activity, capturing not just code but contributor identity, institutional affiliation, and geographic origin. We constructed a large-scale data pipeline to collect and process repository- and contributor-level records spanning 2008–2025. This enabled us to identify repositories where developers from multiple countries co-contributed.

The resulting tabulations and network-based visualizations reveal collaboration patterns across the top contributing nations globally. Beyond measurement, this work surfaces questions of representation: whose contributions are visible in global OSS, which sectors and geographies lead, and where the data goes silent. These questions sit at the heart of what it means to create truly global connections - and why diverse perspectives in data science matter how we build and interpret such indicators.

Multimodal AI Fusion for Better Alzheimer Detection

Chipo Zidana, ZIMYAS

Background: Diagnosing Alzheimer's disease (AD) and its precursor, Mild Cognitive Impairment (MCI), is hindered by hidden correlations between genetic factors and neurostructural changes. Isolated modality evaluation misses critical systemic insights.

Methods: We propose a dual-branch multimodal AI fusion framework integrating functional fMRI and genomic data from

the ADNI database. Significant differentially expressed genes (DEGs) identified via LIMMA in R were mapped to structural brain regions of interest (ROIs) utilizing the UK Biobank. Volumetric spatial and high-dimensional time-series features were extracted using 3D convolutions and tsfresh. Classification was evaluated across Random Forest, SVM, and deep Neural Network architectures.

Results: Our framework identified 12,214 significant probe sets mapping to 10 unique genes regulatory of structural variations in key ROIs like the lateral occipital cortex. Targeted biological ROI feature selection drastically minimized computational noise, outperforming all-feature models. The Neural Network branch achieved a hold-out accuracy and AUC of 1.00, while SVM and Random Forest reached cross-validation accuracies of 97% and 93% respectively.

Conclusion: Fusing genomic variables with neuroimaging eliminates noise and accelerates MCI-to-AD classification accuracy. This interpretable pipeline establishes a powerful paradigm for target-driven precision diagnostics.

End-of-Life Healthcare Utilization among Medicare Beneficiaries with Parkinson's Disease Dementia at USA Academic Movement Disorders Centers: A 2019 Pre-Intervention Assessment

Adriana Perez, The University of Texas Health Science Center at Houston

Evaluating end-of-life (EOL) care for hospitalized Medicare decedents with Parkinson's disease dementia (PDD), defined here as the last 90 days of life in 2019, poses statistical challenges when analyzing cohorts with PDD beneficiaries clustered within hospitals. The four primary outcomes were: non-elective inpatient hospitalization, intensive care unit (ICU) admission, in-hospital death, or receiving palliative care consultation. PDD beneficiaries first visited either a comparison Academic Movement Disorders Center (AMDC) or an AMDC intervention-scheduled center. The adjusted odds ratios (AOR) for the simultaneous effect of visiting AMDCs, controlling for sex, race, age, Social Vulnerability Index (SVI), and the cluster AMDC effect, are reported. The sample sizes were 675, 628, 507, and 486 for each outcome, respectively. There were no statistically significant differences in any health outcomes between PDD beneficiaries who first visited an AMDC intervention-scheduled center and those who first visited an AMDC comparison center. After adjustment, PDD beneficiaries of races other than non-Hispanic white had higher odds of non-elective inpatient hospitalization (AOR=1.92; 95% CI 1.05–3.52), as well as ICU admission (AOR=2.43; 95% CI 1.29–4.58). How AMDC centers and socio-demographic characteristics influence PDD outcomes 90 days before death helps to inform future program design and policy recommendations aimed at optimizing care for PDD Medicare beneficiaries.



Session Info



CONTRIBUTED SESSION

October 6th, 14:00 - 15:00 UTC

Selected Topics in Statistics

USE SCHED TO ACCESS THE ZOOM ROOM

A Statistical Assessment of Urban–Rural Differences in HIV Outcomes Following ART Introduction in Kenya, 2003–2008 | Leveraging Forum Theatre to Enhance Communication and Confidence in Biostatistical Practice | End of Life Healthcare Utilization among Medicare Beneficiaries with Parkinson’s Disease at US Academic Movement Disorders Centers: A 2019 Pre-Intervention Assessment | Motivating young adults to pursue (and hold) the careers of their choice

S503



ORGANIZER: IDWSDS Program Committee

CHAIR: Gaoqianxue Liu, Drexel University

Speaker Bios

S503



MISS TANYA SUN
UC Berkeley

Tanya Sun is a student at the University of California, Berkeley, pursuing a BA in statistics and applied mathematics. Her research focuses on health disparities, infectious disease epidemiology, and the application of statistical methods to population health questions. Tanya is interested in pursuing graduate study in biostatistics and hopes to develop quantitative methods that improve health outcomes in underserved populations.



DR. ADRIANA PEREZ
The University of Texas Health Science Center at Houston

<https://sph.uth.edu/faculty/?fac=86IAK4QbEovkCDrymIVg5g==>

Professor at the Department of Biostatistics and Data Science. Dr. Pérez is engaged in a wide range of research projects: theoretical model evaluation accounting for imputation uncertainty; fitting complex data; analysis of cluster-randomized community trials; clinical trials; analysis of food intake involving measurement error; tobacco regulatory science; and implementation science. Dr. Pérez has expertise in conducting and disseminating research findings from youth and adults, as well as developing new statistical methodologies. She has completed and disseminated the results of several NIH grants. Dr. Pérez has over 150 peer-reviewed publications. Dr. Pérez is an active member of the International Biometric Society (IBS) and the American Statistical Association (ASA). She was the elected president of the Central American and Caribbean Region (RCAC) of IBS. Currently, Dr. Pérez serves as the representative council for RCAC IBS and chair of Inter-Region/Network Funding Task Force.



MS. AMANDA YOUNG
Geisinger

Amanda Young, MS, is Director of the Biostatistics Core at Geisinger, where she leads a team of MS- and PhD-level biostatisticians supporting clinical, epidemiologic, and translational research across the clinical enterprise and the School of Health Sciences. She brings deep expertise in electronic health record data, study design, advanced statistical methods, and real-world evidence, partnering with investigators across specialties to produce rigorous, impactful science. Amanda is also a dedicated educator and mentor, teaching graduate-level courses in biostatistics, epidemiology, and research methods while fostering the development of resident research skills and supporting the next generation of genetic counselors. She is passionate about building collaborative study teams and enabling high-quality, meaningful research.



PROF. HARITINI TSANGARI
University of Nicosia

<https://www.unic.ac.cy/tsangari-haritini/>

Haritini Tsangari is a Professor in the School of Business at the University of Nicosia, Cyprus. She has a BSc in Mathematics and Statistics (University of Cyprus), an MSc and a PhD in Statistics (Pennsylvania State University, USA). She has published more than 150 high-impact journal articles, book chapters and conference proceedings, which are widely cited. Many of her articles have been awarded or selected as Editor's Choice. She has been the principal investigator in projects funded by European funds or national Institutions. Her research involves Applied Statistics in Finance, Management and Health Sciences. Some of her current topics are related to job satisfaction, gender equality, AI in daily life and workplace inclusion for people with disabilities. Prof. Tsangari is the Vice President of the Cyprus Statistical Society.



A Statistical Assessment of Urban–Rural Differences in HIV Outcomes Following ART Introduction in Kenya, 2003–2008

Tanya Sun, UC Berkeley

The purpose of this study was to examine the effects of antiretroviral therapy (ART) upon introduction to Kenya in 2003–2008, focusing on the urban–rural difference in treatment access and its implications for HIV-related outcomes. Previous studies (Dos Santos et al., 2025) suggest rural communities suffer from isolation to medical facilities, resulting in lack of adherence to HIV maintenance care and poor health outcomes. The age distributions of HIV-positive Kenyan women (in 2003 and 2008) were analyzed separately for urban and rural regions according to Demographic and Health Surveys (DHS) classification. To examine these effects over time, statistical testing (t-tests, p-values, regression models) was used to evaluate significant change in age distributions. Findings indicated that the implementation of ART increased the average age of HIV-positive women by 1.8 years, with negligible difference between regional settings. However, when the relationship was examined with respect to wealth inequality, the highest quintiles of HIV-positive women in urban areas displayed significant increases in age. These findings suggest that, despite greater proximity to medical facilities in urban areas, this nearness does not necessarily translate into improved medical access due to social inequality. Thus, study results highlight the need for efforts to reduce health disparity within urban regions, and further study into the contribution of regional wealth inequality may be of interest.

Leveraging Forum Theatre to Enhance Communication and Confidence in Biostatistical Practice

Amanda Young, Geisinger

Biostatisticians in research cores within learning health systems face unique interprofessional communication and collaboration (ICC) challenges when working with clinicians and learners from diverse backgrounds. Traditional training often lacks experiential, emotionally engaged practice needed to build confidence in these situations. To address this, we developed a Forum Theatre–based professional development program for biostatisticians.

Forum Theatre, created by Augusto Boal, uses short scenes of unresolved conflict and invites participants (“spect-actors”) to intervene, replace the protagonist, and rehearse alternative approaches. In this workshop, participants explored common scenarios: navigating ethical or methodological conflicts, managing imposter syndrome, and advocating for recognition. Sessions combined staged scenes, structured audience analysis, small-group strategy development, and interactive reenactments.

Participants reported that the format provided a psychologically safe yet realistic space to practice communication, experiment with language, and build a shared vocabulary around ICC challenges. The approach also reinforced that these challenges are widely shared. By engaging emotional and relational dynamics, this model offers a practical framework for strengthening communication skills among biostatisticians and other research professionals.

End of Life Healthcare Utilization among Medicare Beneficiaries with Parkinson’s Disease at US Academic Movement Disorders Centers: A 2019 Pre-Intervention Assessment

Adriana Perez, The University of Texas Health Science Center at Houston

Evaluating end-of-life (EOL) care for hospitalized Medicare decedents with Parkinson’s disease (PD), defined here as the last 90 days of life in 2019, poses statistical challenges. The four primary outcomes were: non-elective inpatient hospitalization, intensive care unit (ICU) admission, in-hospital death, or receiving palliative care consultation. Beneficiaries first visited either a comparison Academic Movement Disorders Center (AMDC) or an AMDC intervention-scheduled center. The adjusted odds ratios (AOR) for the simultaneous effect of visiting AMDCs, controlling for sex, race, age, Social Vulnerability Index (SVI), and the cluster AMDC effect, are reported. The sample sizes were 2040, 2029, 1984, and 1819 for each outcome, respectively. There were statistically significant differences in ICU admission between beneficiaries first visiting an AMDC intervention-scheduled center versus an AMDC comparison center (AOR=1.27; 95% CI 1.04–1.55). Quartiles 3 (AOR=1.57; 95% CI 1.03–2.41) and 4 (AOR=1.72; 95% CI 1.07–2.75) of the SVI were associated with in-hospital death. After adjustment, beneficiaries of races other than non-Hispanic white had higher odds of palliative care consultation than non-Hispanic white beneficiaries (AOR=1.91; 95% CI 1.12–3.26). How AMDC centers and socio-demographic characteristics influence PD outcomes 90 days before death helps to inform future program design and policy recommendations aimed at optimizing care for PD Medicare beneficiaries.

Motivating young adults to pursue (and hold) the careers of their choice

Haritini Tsangari, University of Nicosia

We live in the era of AI and digital transformation, with rapid technological changes. Children who are now in primary school are expected to work in the future in specializations and professions that do not even exist today. Organizations need to explore new ways to increase productivity and business performance. Matching employees’ skills with the job market appears to be a challenge, especially when considering young adults. This special group of employees, who are recent university graduates, is evidenced to be constantly shifting between jobs until expectations are met. The presentation focuses on identifying practices and conditions under which young employees will experience high levels of work motivation and job satisfaction. A comprehensive modeling framework has been developed, testing the effect of “job dimensions”, such skill variety, job significance and feedback, technology use, decision-making and autonomy on “personal and work outcomes”. This relation is mediated by employees’ psychological states, such as self-confidence and experienced meaningfulness of the work. The discussion stresses the importance of effective job design and redesign strategies that are focused on young employees, aiming to match their qualifications and skills with the careers of their choice.



Session Info

S504

TECHNICAL STATISTICAL SESSION

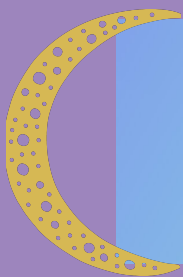
CONTRIBUTED SESSION

October 6th, 15:00 - 16:00 UTC

Additional Topics in Health and Biostatistics

USE SCHED TO ACCESS THE ZOOM ROOM

Correcting Bias from Covariate-Dependent Censoring in Survival Disparity Decomposition | End-of-Life Healthcare Utilization among Medicare Beneficiaries with Parkinson's Disease- Alzheimer's Disease-Related Dementia at USA Academic Movement Disorders Centers: A 2019 Pre-Intervention Assessment | End-of-Life Healthcare Utilization among Medicare Beneficiaries with Parkinson's Disease Dementia at USA Academic Movement Disorders Centers: A 2019 Pre-Intervention Assessment | Physical Activity Patterns by Age Among U.S. Adults: Associations with Chronic Conditions, Health Behaviors, and Health-Related Quality of Life

ORGANIZER: IDWSDS Program Committee**CHAIR:** Daniela Márquez, Universidad Autónoma de Chiapas

Speaker Bios

S504



DR. ADRIANA PEREZ

The University of Texas Health Science Center at Houston

<https://sph.uth.edu/faculty/?fac=86IAK4QbEovkCDrymIVg5g==>

Professor at the Department of Biostatistics and Data Science. Dr. Pérez is engaged in a wide range of research projects: theoretical model evaluation accounting for imputation uncertainty; fitting complex data; analysis of cluster-randomized community trials; clinical trials; analysis of food intake involving measurement error; tobacco regulatory science; and implementation science. Dr. Pérez has expertise in conducting and disseminating research findings from youth and adults, as well as developing new statistical methodologies. She has completed and disseminated the results of several NIH grants. Dr. Pérez has over 150 peer-reviewed publications. Dr. Pérez is an active member of the International Biometric Society (IBS) and the American Statistical Association (ASA). She was the elected president of the Central American and Caribbean Region (RCAC) of IBS. Currently, Dr. Pérez serves as the representative council for RCAC IBS and chair of Inter-Region/Network Funding Task Force.



MS. GLORIA OPPONG

University of Tennessee at Chattanooga

<https://www.linkedin.com/in/gloria-oppong-910610324/>

Gloria Oppong is an incoming Ph.D. student in Biostatistics at Florida State University. She recently earned her M.S. in Mathematics with a concentration in Applied Statistics from the University of Tennessee at Chattanooga. Her research focuses on public health statistics, epidemiology, and statistical modeling using large-scale health survey data. She has conducted research using the CDC Behavioral Risk Factor Surveillance System (BRFSS), examining physical activity, chronic disease, and health-related quality of life among U.S. adults. Gloria has also contributed to research in quantum computing and combinatorial optimization. She is passionate about applying statistical methods to address public health challenges and improve health outcomes through data-driven research.



MRS. SABRINA ZHANG

Westat

Sabrina Zhang has been working as a survey statistician at the statistics and data science department at Westat for over 10 years. At Westat, she works on sampling, weighting, propensity models, data linkage, and other topics across education, health, transportation surveys and so on. She is currently pursuing her PhD degree at Joint Program of Survey Methodology at University of Maryland. She has also been a long-term member of ASA and WSS and currently serves as a Quantitative Literacy Committee Chair at WSS.



Correcting Bias from Covariate-Dependent Censoring in Survival Disparity Decomposition*Sabrina Zhang, Westat*

Research on racial and socioeconomic health disparities (HD) in time-to-event outcomes is crucial to public health. Yet, existing HD decomposition methods can yield biased estimates when censoring, such as dropout or competing risks, depends on covariates, including demographic, socioeconomic, or behavioral characteristics. Widely used Peters-Belson approaches combine Kaplan-Meier estimation with Cox regression, but these methods rely on different censoring assumptions that may not hold under covariate-dependent censoring. We propose a novel decomposition method that integrates propensity score (PS) methodology with inverse probability of censoring weighting (IPCW). The PS component balances covariates across groups to estimate counterfactual survival without specifying an outcome model, while IPCW corrects for dependent censoring by reweighting uncensored individuals to represent those who were censored. Within a semiparametric framework, the proposed IPCW-PS approach reduces bias compared with conventional Kaplan-Meier/Cox methods. Simulation studies demonstrate substantial bias reduction under dependent censoring. We apply the method to NielsenIQ Consumer Panel data to examine smoking cessation disparities between White and Black households and find that smoke-free laws explain a meaningful portion of the observed disparity.

End-of-Life Healthcare Utilization among Medicare Beneficiaries with Parkinson's Disease- Alzheimer's Disease-Related Dementia at USA Academic Movement Disorders Centers: A 2019 Pre-Intervention Assessment*Adriana Perez, The University of Texas Health Science Center at Houston*

Evaluating end-of-life (EOL) care for hospitalized Medicare decedents with Parkinson's Disease- Alzheimer's Disease-Related Dementia (PD-ADRD), defined here as the last 90 days of life in 2019, poses statistical challenges. The four primary outcomes were: (i) non-elective inpatient hospitalization, (ii) intensive care unit (ICU) admission, (iii) in-hospital death, or (iv) receiving palliative care consultation. Beneficiaries first visited either a comparison Academic Movement Disorders Center (AMDC) or an AMDC intervention-scheduled center. The adjusted odds ratios (AOR) for the simultaneous effect of visiting AMDCs, controlling for sex, race, age, Social Vulnerability Index (SVI), and the cluster AMDC effect, are reported. The sample sizes were 1362, 1344, 1230, and 1100 for each outcome, respectively. There were no statistically significant differences in (i) non-elective inpatient hospitalization (AOR=0.41; 95% CI 0.12–1.36), (ii) ICU admission (AOR=0.80; 95% CI 0.14–4.68), (iii) in-hospital death (AOR=0.68; 95% CI 0.07–7.23) and (iv) receiving palliative care consultation (AOR=0.50; 95% CI 0.05–4.93) between beneficiaries who first visited an AMDC intervention-scheduled center and those who first visited an AMDC comparison center. How AMDC centers and socio-demographic characteristics influence PD-ADRD outcomes 90 days before death helps to inform future program design and policy recommendations aimed at optimizing care for PD-ADRD Medicare beneficiaries.

Physical Activity Patterns by Age Among U.S. Adults: Associations with Chronic Conditions, Health Behaviors, and Health-Related Quality of Life*Gloria Oppong, University of Tennessee at Chattanooga*

Introduction: Physical activity (PA) is essential for healthy aging, yet PA mode, frequency, duration, and intensity vary across adulthood and relate to health-related quality of life (HRQOL), chronic conditions, and health behaviors. This study examined age-stratified PA patterns and their associations with selected health indicators.

Methods: Data from 433,323 adults in the 2023 Behavioral Risk Factor Surveillance System (BRFSS) were analyzed in this cross-sectional study. Participants were grouped as younger (18–44 years), middle-aged (45–64 years), and older (≥65 years). Logistic regression estimated odds ratios (ORs, 95% CI) for not meeting aerobic and muscle-strengthening Physical Activity Guidelines (PAGA), adjusting for sociodemographic and health-related covariates.

Results: Overall, 75.3% reported PA in the past month; 61.9% met aerobic PAGA and 40.1% met muscle-strengthening PAGA. Walking increased with age (44.4%, 58.3%, and 59.4%), while gardening and household chores increased and weightlifting and running/jogging declined. Obesity, smoking, diabetes, asthma, and female sex were associated with higher odds of not meeting aerobic PAGA, particularly among older adults.

Conclusion: PA patterns shift across adulthood, supporting age-tailored interventions promoting aerobic and muscle-strengthening activity, especially among older adults with multiple risk factors.



Session Info

S505

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 18:00 - 19:00 UTC

Code is for Everyone: Showcasing Tools to Support Learners at all Levels

USE SCHED TO ACCESS THE ZOOM ROOM

Writing code can feel like a challenging task, but it doesn't have to be one! In this session, sponsored by the Caucus for Women in Statistics and Data Science, we feature a diverse group of speakers with experience developing tools that simplify code-related tasks from documentation and debugging to deliverables. First, Dr. Kelly Bodwin will present two Quarto extensions that allow the user to highlight renderable code chunks and separate a document into multiple source files, providing tips for Quarto extension development along the way. Next, Dr. Mine Çetinkaya-Rundel will discuss an R package that leverages large language models to provide immediate feedback on student work, highlighting both technical and ethical considerations. Finally, Dr. Cara Thompson will conclude with a live demonstration of how you can use the ggplot2 package in R to create customized, reusable functions for producing accessible, well-designed graphics. The three speakers will cover a wide array of tools, development tips, and use cases, making the session interesting to a general audience. However, all three showcase how strategic deployment of technical tools can support all members of the global statistics community, no matter their comfort level with code!

ORGANIZER: Ashley Mullan, Vanderbilt University**CHAIR:** Jamie Joseph, Henry Ford Health

Speaker Bios

S505



DR. KELLY BODWIN

California Polytechnic State University

Dr. Kelly Bodwin is an Associate Professor of Statistics and Data Science at California Polytechnic State University, San Luis Obispo. Her research interests include cross-disciplinary work in Digital Humanities, including historical social networks and authorship identification; methodologies for high-dimensional clustering and cluster validation; and Data Science education especially as it pertains to machine learning and software. She is the author of several open-source tools (flair, tidyclus, flourish, versions) as well as public educational materials for Data Science with R and python.



DR. MINE CETINKAYA-RUNDEL

Duke University

Dr. Mine Çetinkaya-Rundel is Professor of the Practice and the Director of Undergraduate Studies at the Department of Statistical Science and the Director of First-Year Experience at Duke University. She is also a Developer Educator at Posit, PBC, an open-source data science software company, formerly RStudio Inc. Mine has co-authored four open-source statistics textbooks, is the creator and maintainer of Data Science in a Box, co-leads the international ASA DataFest, and has been developing and teaching various massive open online courses, including the popular Statistics with R specialization on Coursera. In 2021, Mine was awarded the Robert V. Hogg Award for Excellence in Teaching Introductory Statistics. Her work focuses on innovation in statistics and data science pedagogy, with an emphasis on computing, reproducible research, student-centered learning, and open-source education.



DR. CARA THOMPSON

Building Stories with Data LTD

Dr. Cara Thompson is a data visualisation consultant with an academic background, specialising in helping research teams and data-driven organisations turn their data insights into clear and compelling visualisations.

After her PhD in Psychology and a spell teaching research methods at Edinburgh Uni, she embarked on a career in psychometrics at the Royal college of Surgeons of Edinburgh. After ten years of helping surgeons and other medical professionals understand complex patterns in exam data, she set out as an independent data visualisation consultant, to continue building creative solutions for a range of different organisations.

She lives in Edinburgh, Scotland, with her husband and two young daughters. Cara regularly shares coding tips for dataviz online and genuinely enjoys helping others level up their dataviz skills through talks, bespoke toolkits and organisational training.



Adaptive Functioning in Angelman Syndrome: Insights from Statistical Modeling*Julia Zhang, RTI International*

Modeling person ability scores in rare disease natural history studies presents several methodological challenges, including nonlinear developmental trajectories, heterogeneous variance, and floor and ceiling effects. This presentation describes a modeling framework developed to address these challenges using data from a large, multi-site longitudinal study of individuals with Angelman syndrome (AS).

AS is a rare neurogenetic condition characterized by severe to profound intellectual disability, minimal speech, and motor impairments. We focus on modeling adaptive functioning over age, using Vineland-3 Growth Scale Value scores across multiple subdomains. To capture the complex structure of these data, we evaluated and compared a series of candidate models and selected an approach that balances interpretability with statistical rigor while accommodating key features of the data, including nonlinearity and heteroscedasticity.

We present the model development process, including comparison of alternative specifications, and highlight how the selected framework captures distinct developmental trajectories across molecular subtypes of AS. Results demonstrate meaningful differences in patterns of adaptive functioning by subtype and illustrate how model choice influences substantive conclusions. We conclude with practical recommendations for modeling longitudinal person ability scores in rare disease contexts, with implications for study design, analysis, and interpretation.

Novel Influence Diagnostics in Multistate Models for Breast Cancer*Valeria Leiva, Department of Public Health, Universidad de los Andes, Chile*

Multistate models were developed to model survival data where several midpoints and endpoints are of interest; and they have been particularly successful in modeling dynamics of cancer. As in any statistical model, identification of influential observations is an essential task, as they can significantly affect the validity of inferred parameters and conclusions drawn from the data. The local influence approach is a set of methods designed to detect the effect of small perturbations of the model or data on the inference, allowing for deeper data analysis. In this paper, we derive local influence methods for multistate models and illustrate their use with a breast cancer dataset. In particular, we develop and implement local influence diagnostic techniques based on a suitable estimation equation. For simplicity, we restrict our consideration to the Multistate Proportional Hazards model, using different case-weight perturbation strategies.

lacks a rigorous goodness-of-fit (GoF) framework to validate its assumptions. Existing methodologies face computational barriers and lack documented performance for dimensions $d > 2$. This study evaluates the implementation, test size, and Power of standard Empirical Distribution Function (EDF) tests, specifically Kolmogorov-Smirnov (KS), Anderson-Darling (AD), and Cramér-von Mises (CvM), alongside a novel Parametric Bootstrap Chi-square (PB-CS) test. Implementing EDF tests for the Dirichlet distribution introduces two major hurdles: the absence of a closed-form distribution function for $d > 2$ and the mathematical ambiguity of ranking multivariate observations. To bypass the closed-form limitation, we propose a two-step transformation using Beta decomposition and the Rosenblatt transformation to map $(d-1)$ -dimensional unit simplex data into a $(d-1)$ -dimensional unit hypercube. Furthermore, we demonstrate that the frequency-based PB-CS test completely bypasses the multivariate ranking problem. To adapt it for continuous data, we introduce an extended Doane's rule for bin selection paired with a quasirandom binning strategy using Sobol sequences and Voronoi partitioning. Extensive simulations show that while EDF methods thrive in lower dimensions, the PB-CS framework delivers superior or comparable performance as dimensionality increases.

A Bayesian Exploration of Bimodal Fertility Patterns Using Lifetime Densities.*Shambhavi Singh, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India.*

The work explores mixtures of three lifetime densities, viz., normal, Weibull, and generalized gamma densities, to capture fertility patterns characterized by hump- or bimodal-shaped behavior, an intriguing trend that researchers have observed over the last two to three decades. It harnesses the Bayesian framework to achieve a comprehensive posterior analysis. The Metropolis algorithm was used with a carefully selected candidate-generating density to produce posterior samples for each case. Numerical illustrations have been conducted using unabridged real-life data. Ultimately, the proposed models were compared to identify the most suitable model, ensuring compatibility with the available data. This work is inspired by previous studies that have largely overlooked the use of more flexible lifetime models in analyzing fertility patterns of this nature, prompting us to incorporate them into our analysis.

Goodness-of-fit test for the Dirichlet distribution*Sucharitha Dodamgodage, Clarkson University*

The Dirichlet distribution is the first choice for modeling compositional data like microbiome profiles, yet the literature



Session Info

S506

TECHNICAL STATISTICAL SESSION

INVITED SESSION

October 6th, 20:00 - 21:00 UTC

Recent Advances in Sequential Inference

USE SCHED TO ACCESS THE ZOOM ROOM

Sequential analysis continues to play a central role in modern statistics by enabling data-driven decision making while balancing statistical efficiency, computational cost, and practical constraints. This session highlights recent methodological advances in sequential and multistage inference across a range of contemporary statistical problems. Topics include sequential estimation of distributional overlap with boundary corrections for reliable confidence interval construction, multistage algorithms for efficient data compression under growing dimensions, penalized sequential estimation and variable selection for recurrent event models with high-dimensional covariates, and multistage sampling strategies for constructing minimum-risk confidence regions for multivariate means. Together, these contributions illustrate how sequential methodologies are being extended beyond their classical foundations to address challenges arising from high-dimensional data, adaptive sampling, optimization of statistical risk, and modern computational applications.

ORGANIZER: Swathi Venkatesan, Fairfield University**CHAIR:** Swathi Venkatesan, Fairfield University

Speaker Bios

S506



DR. SWARNALI BANERJEE
Loyola University Chicago

Dr. Swarnali Banerjee is an Associate Professor and Director of Data Science in the Department of Mathematics and Statistics at Loyola University Chicago. She received her Ph.D. in Statistics from the University of Connecticut in 2014. Prior to joining Loyola University in 2016, she served as a tenure-track Assistant Professor in the Department of Mathematics and Statistics at Old Dominion University from 2014 to 2016. She earned her Bachelor's degree in Statistics from the University of Calcutta and her M.Stat. degree from the Indian Statistical Institute. Her research interests include sequential analysis, statistical methods for bioinformatics, the analysis of count data arising in ecology and the health sciences, and multistage methodologies for modern data applications.



DR. ZHE WANG
Denison University

Dr. Zhe Wang is an Assistant Professor in the Department of Data Analytics at Denison University. She earned her B.S. in Statistics from Beijing Normal University and her M.S. and Ph.D. in Statistics from the University of Connecticut. Her research interests include statistical inference, sequential methods, and interdisciplinary applications of statistics, particularly in healthcare and biomedical sciences. Her work combines rigorous mathematical modeling with practical data analysis to develop innovative methodologies for complex real-world problems. She is passionate about the role of uncertainty in statistical inference and enjoys collaborating across disciplines to advance data-driven research.



DR. SWATHI VENKATESAN
Fairfield University

Dr. Swathi Venkatesan is an Assistant Professor in the Department of Mathematics at Fairfield University. She earned her PhD in Statistics from the University of Connecticut. Her research interests include sequential estimation and inference, experimental design, and the development of statistical methodologies for efficient and reliable data analysis. Her work focuses on advancing theoretical and applied methods in statistics.



DR. LAURA DUMITRESCU
Fairfield University

Dr. Laura Dumitrescu is an Associate Professor in the Department of Mathematics at Fairfield University, where she has been a faculty member since 2022. Prior to joining Fairfield, she was a member of the Probability and Statistics research group in the School of Mathematics and Statistics at Victoria University of Wellington, New Zealand, where she served as a Lecturer beginning in 2015. Before that, she was a researcher at CREST, Laboratoire de Statistique et Modélisation, ENSAI, France. Dr. Dumitrescu earned her Ph.D. in Mathematics from the University of Ottawa, Canada, in 2011. She subsequently held a Postdoctoral Fellowship at Carleton University and a Visiting Professorship in the Department of Mathematics at Washington University in St. Louis. Her research interests include mathematical statistics, asymptotic inference, martingale methods in statistics, nonparametric statistics, biostatistics, high-dimensional dependent data, and the theory of sample surveys.





Novel Compression Algorithms Using Multi-Stage Methods

Swarnali Banerjee, Loyola University Chicago

Our growing research demands bigger datasets which also makes compression algorithms the need of the hour. Since most of the data sizes are dynamic, sequential and multi-stage algorithms are very useful to construct such algorithms. For the past few years our research has focused on constructing multi-stage methods for data compression algorithms using Principal Component Analysis. For data arriving sequentially, however, several online PCA methodologies exist which focus only on compression risk minimization. Our work focusses on minimizing the cost-compression risk which considers the compression loss and the sampling cost using a multistage PCA procedures. While our previously published work focusses on fixed dimensions, our current work explores these algorithms under “growing dimensions”. Under certain regularity conditions, these procedures are shown to be first-order and second order efficient and risk efficient. The proposed algorithms are novel, fast and efficient on various types of data as will be seen in illustrations.

Sequential Estimation and Boundary Correction for the Overlap Coefficient

Zhe Wang, Denison University

Overlap quantifies the shared area between two probability distributions and serves as an intuitive measure of similarity between populations. The Overlap Coefficient (OVL) provides a numerical value for this shared region, ranging from 0 to 1. In this presentation, we develop both fixed-sample and sequential estimation methodologies for constructing fixed-size confidence intervals of the OVL. Special attention is given to challenges that arise when the true OVL lies near the boundaries (0 or 1), where traditional normal approximations often fail. We discuss variance-stabilizing transformations and boundary corrections for improving the performance of sequential stopping rules. Simulation studies demonstrate the efficiency and accuracy of the proposed methods across small, moderate, and large OVL settings, followed by applications to real-world data examples.

Penalized Sequential Estimation for Recurrent Events with High-Dimensional Covariates

Laura Dumitrescu, Fairfield University

The study of recurrent events, such as hospital readmissions or stock market volatility spikes, is important in biostatistics and finance but poses statistical challenges due to complex dependencies and censoring. Under a semiparametric framework for modeling gap times between events, we employ a penalized sequential estimating equations approach to handle high-dimensional covariates. The procedure yields estimators with attractive large sample properties and we illustrate how the rate of divergence in the number of parameters affects the convergence rate of penalized estimators when the number of covariates grows to infinity as the number of clusters increases. The proposed method has the advantage of simultaneously estimating parameters while selecting important variables and accounting for the within-cluster correlation.

Sequential Sampling for Ellipsoidal Confidence Regions for Multivariate Means

Swathi Venkatesan, Fairfield University

Traditional fixed-size confidence region (FSCR) methods for estimating the mean of a multivariate normal distribution often fix the region's maximum diameter in advance, without regard to the quality of available data. We propose a new approach that incorporates data quality into determining the region's size. Starting with a modified FSCR method where the structure of the variance-covariance matrix is known, we introduce a minimum risk FSCR (MRFSCR) framework inspired by point estimation methods that balance estimation accuracy and sampling cost. We develop a unified multistage sampling strategy to construct these regions, ensuring desirable asymptotic properties. The methodology is illustrated through practical sampling strategies, simulation studies, and real-data examples.



Session Info



S507

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

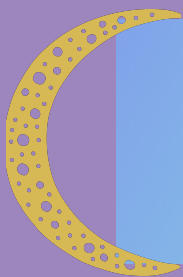
INVITED SESSION

October 6th, 21:00 - 22:00 UTC

CWS Special Session for Award Winners

USE SCHED TO ACCESS THE ZOOM ROOM

This session will have two award winners from the Caucus for Women in Statistics & Data Science. The Woodroffe Award winner, Xiaojing Wang, and the Societal Impact Award winner, Claire Kelling, will share about their work and achievements.

ORGANIZER: Caucus for Women in Statistics and Data Science**CHAIR:** Dong-Yun Kim, NHLBI/NIH

Speaker Bios

S507

**DR. CLAIRE KELLING**

Carleton College

Claire Kelling is an Assistant Professor of Statistics at Carleton College who earned her Dual PhD in Statistics and Social Data Analytics at Penn State. She specializes in working with community organizations/members, government agencies, faculty, and students to answer complex questions related to public policy with data and statistical methods. Her primary focus recently is on the development of statistical methods in spatial statistics for the analysis of policing data, driven by questions from and partnership with community members.

**PROF. XIAOJING WANG**

University of Connecticut

Xiaojing Wang is Professor of Statistics at the University of Connecticut. Her research centers on Bayesian latent variable models and scalable inferential methods for complex, high-dimensional, and multi-source data, with applications in educational measurement, personalized medicine, ecology, neuroscience, and mobile health. She is the recipient of the 2026 Michael Woodroffe Award, the NSF CAREER Award, the UConn Expanding Research Impact Award, and the 2026 ICSA President's Citation Award. She is an NIH mHealth Training Institute Scholar and an elected member of the International Statistical Institute. She serves as Associate Editor of the Journal of Computational and Graphical Statistics, Psychometrika, and Sankhya A; Co-Chair of the ICSA Program Committee; Chair-Elect of the ASA Social Statistics Section; and Chair of the ASA Statistical Partnerships Among Academe, Industry, and Government Committee.





When Is More Data Worth It? Quantifying the Value of Auxiliary Information in Bayesian Latent Variable Models

Xiaojing Wang, University of Connecticut

Modern assessments provide diverse sources of information—including item responses, response times, and external test scores—that may improve estimation of a latent construct of interest. Joint modeling can integrate these sources, but how much does each auxiliary source contribute, and when is incorporating additional data beneficial? Existing model assessment methods evaluate the joint model as a whole, providing limited insight into the incremental value of individual sources. We develop a general Bayesian framework for quantifying the added value of auxiliary information in latent variable models. We derive component-wise decompositions of commonly used Bayesian model assessment criteria and introduce novel Δ -diagnostics that quantify the incremental gain from each auxiliary source in estimating the latent construct of interest. To enable stable and scalable Bayesian computation, we reparameterize the joint model using a decoupled latent architecture. We demonstrate the framework using AppRISE, a gamified early-literacy screener designed to identify learning risk among children aged 4–8 and support timely intervention. The framework extends naturally to cognitive diagnosis models and applies broadly to mobile health, neuroscience, and other settings in which multiple data sources are integrated to characterize latent constructs of interest.

My Why, Where, and How: Lessons from a Community-Engaged Statistics Research Partnership

Claire Kelling, Carleton College

Some of the most fulfilling and challenging experiences of my career have come from being part of community-engaged research teams. In this talk, I'll share why I chose this path, introduce the neighborhoods and community partners I work with, and talk about the critical expertise that these local residents bring to our work. Community-engaged research can present avoidable obstacles, such as mismatched expectations around dissemination and timing. I'll walk through the structures and mindset that helped me set partnerships up for success from the beginning. I'll close with practical advice for anyone considering this kind of work, along with resources for getting started.



Session Info

S508

TECHNICAL STATISTICAL SESSION

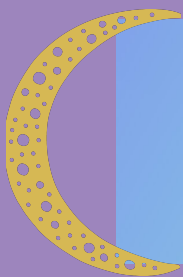
INVITED SESSION

October 6th, 22:00 - 23:00 UTC

Statistical Methods and Applications in One Health

USE SCHED TO ACCESS THE ZOOM ROOM

This session will showcase statistical methods and data-driven approaches for One Health, reflecting the interconnected health of humans, animals, and the environment. Presentations by women graduate students in statistics will highlight innovative methodological and applied research across One Health domains, spanning the development and application of statistical methods for handling complex, high-dimensional, and heterogeneous data, improving prediction and inference, and addressing uncertainty in human, animal, and environmental health systems. The session also provides a platform for emerging researchers to share their work, engage with the statistical community, and foster collaboration and mentorship.

ORGANIZER: Elif Acar, University of Guelph**CHAIR:** Elif Acar, University of Guelph

Speaker Bios

S508



MS. SONJA FRIESEN
University of Guelph

Sonja graduated from the University of Manitoba with a BSc in Data Science with a minor in Biological Sciences and recently began a Master's in Statistics with a One Health Concentration at the University of Guelph. She enjoys working on any project that combines health care, R programming, and data science.



MS. AYISHA COK
University of Guelph

Ayisha N. COK is a Ph.D. student in Statistics at the University of Guelph. Her research focuses on developing statistical methods for high-dimensional biomedical data, with interests in measurement error, genotype imputation, meta-analysis, Bayesian methods, and integrative data analysis. Her current work aims to develop bias-correction methods for association studies with applications in genomics, microbiome research, and antimicrobial resistance within the One Health framework.



MS. HESHANI MENDIS
University of Manitoba

Heshani Mendis is an MSc student in Statistics at the University of Manitoba, Winnipeg, Canada. Her research interests lie in statistical machine learning, with a focus on tree-based ensemble methods and flexible splitting criteria for random forests. Her current work introduces a flexible child-node weighting framework for random forest splitting.



MS. NIMSARA DISSANAYAKA
University of Guelph

Nimsara Dissanayaka a PhD student in Applied Statistics at the University of Guelph. Her research focuses on circular data analysis mainly the development of circular regression models, model assessment, and prediction. She is particularly interested in environmental applications of statistical methods for directional data.





Robust inverse normal transformation-based tests under linear mixed effects models

Sonja Friesen, University of Guelph

Many biomedical applications generate vast amounts of data that cannot be managed under human oversight alone, necessitating automated analysis pipelines. These pipelines require robust statistical methods to handle frequent assumption violations in biological settings. Our research is motivated by the International Mouse Phenotyping Consortium (IMPC), which examines mammalian gene function using genetically edited knockout mice. The IMPC employs a linear mixed effects model (LMM) that considers both fixed effects, such as gene, sex, and weight, and random effects that may arise from shared experimental batch or litter. Using simulations based on the IMPC experiment setup, we assess the robustness of the LMM under violations of the normality assumption for error and random effects. We further investigate whether rank-based inverse normal transformations could enhance the detection power of gene effects in these scenarios.

A Closed-Form Correction for Attenuation Bias in Imputed-Genotype Association Studies

Ayisha COK, University of Guelph

Genotype imputation is now standard in genome-wide association studies, yet using imputed genotypes in place of true genotypes in association testing has statistical consequences that are easy to overlook. Whether the imputed genotype is summarised as a dosage or as a best guess, the estimated genetic effect is attenuated toward zero and the residual variance is inflated, so both the effect estimate and its standard error, and hence any inference based on them, are biased when imputed genotypes are analysed directly. In this work, we derive a bias-corrected estimator of the genetic effect and its standard error using imputation quality metrics and genotype variance under Hardy-Weinberg equilibrium, both available from standard imputation workflows. Across a range of effect sizes, allele frequencies, and imputation qualities, our numerical studies show the corrected estimator recovers the true genetic effect, with the attenuation almost entirely removed. Association studies usually focus on power, but our goal is unbiased estimation, so the reported effect reflects the true genetic signal rather than an artefact of imputation. The same attenuation appears wherever an error-prone measurement substitutes for a true quantity. This makes it a single statistical problem shared across the interconnected domains of One Health, from host genetics to the microbiome and environmental surveys.

Flexible Tree Ensembles with Data Adaptive Splitting Rules

Heshani Mendis, University of Manitoba

Accurate split selection in regression trees and random forests is critical for predictive performance. In regression trees, splits are selected by minimizing within-node variances weighted by node proportions. These weights directly influence the impurity reduction criterion. We modify the standard weighting scheme by applying power transformations to node proportions and

examine how this modification influences split selection and predictive accuracy. Based on the proposed weighting framework, we develop four algorithmic variants that introduce flexibility at the node, tree, and forest levels, along with a randomized weight-selection strategy. The proposed framework is evaluated via simulations across a range of data-generating mechanisms, including varying signal strengths, noise levels, functional relationships, and covariate dependence structures and distributions. Performance is assessed using relative prediction error against the default weighting scheme, and the results demonstrate notable improvements, with the tree-level flexible method achieving improved or comparable predictive performance relative to the standard random forest in the majority of scenarios considered. To illustrate practical applicability, the proposed framework is applied to real data, including a Parkinson's disease telemonitoring dataset.

Prediction and Model Assessment of Regression Models with Circular Data

Nimsara Dissanayaka, University of Guelph

Angular or circular data arise in many applications, such as environmental studies, where observations are recorded as directions or angles. Examples include wind direction, animal movement, bird migration, and ocean current direction. In such cases, statistical analyses must properly account for the circular nature of the observations. In this work, we study the consequences of ignoring circularity by comparing the predictive performance of parametric circular regression models with that of traditional linear models in settings with circular responses and/or circular covariates. The results of numerical studies indicate a substantial loss in prediction accuracy when circularity is ignored. We further evaluate parametric circular regression models by comparing them with their nonparametric counterparts under various circular distance metrics and predictive accuracy criteria. The analysis is illustrated using wind direction data to demonstrate the importance of accounting for circularity and its potential applicability to other directional data, such as animal movement.



Sponsors

The 2026 IDWSDS Organizing Committee gratefully acknowledges our sponsors for their generous support of this year's program. Their contributions help make this year's conference possible and allow us to bring together an outstanding program for our attendees.

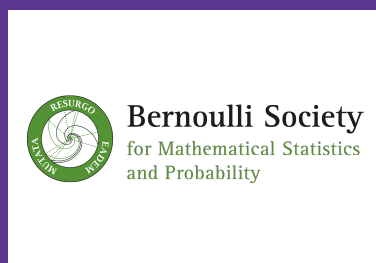
Platinum Level



Gold Level



Silver Level



Bronze Level



Thank you for joining us at this year's conference! We are grateful to everyone who helped make IDWSDS possible. Our organizers, presenters, volunteers, and attendees are at the heart of this community, and your contributions make this celebration of women in statistics and data science so meaningful. Together, we shared a day filled with inspiration, connection, and encouragement. We hope the energy and sense of community we created together continues to carry forward throughout the year. The accomplishments, passion, and global impact of women in statistics and data science continue to grow. Thank you for being part of that momentum, and we look forward to celebrating with you again next year.

