Highlighted Presentations

The 2026 Joint Statistical Meetings will bring together statisticians and data scientists from around the world from Saturday, August 1, to Thursday, August 6. This year, JSM will be held in Boston, Massachussetts. This tip sheet highlights interesting presentations from the conference. Complimentary press registration is open, courtesy of the ASA. Email edoffice@amstat.org for more information.

Highlighted Presentations

MONDAY, AUGUST 3

Developing a Workflow for AI-Based Image Interpretation of Forest Inventory Plots

Forest inventory is painstaking, repetitive work, which is the kind of task AI tools show real promise for. But promise and practice don't always align. Working with the USDA Forest Service's Forest Inventory and Analysis program, the research team put large language models to the test on a concrete task: interpreting forest images. The result was a working prototype, a practical introduction to the ELLMER package in R for integrating LLMs into statistical workflows and consideration of the important question of what quality assurance looks like.

Inference on Deep Neural Networks

While deep neural networks are used for prediction, inference on DNN-estimated subject-specific means for categorical or exponential family outcomes remains underexplored. We address this by proposing a DNN estimator under generalized nonparametric regression models and developing a rigorous inference framework. Through simulations under nonparametric logistic, Poisson, and binomial regression models, we demonstrate the effectiveness and efficiency of our method. We further apply the method to the electronic intensive care unit dataset, a large-scale collection of anonymized health records from ICU patients, to predict ICU readmission risk and offer patient-centric insights for clinical decision-making.

Mining AI for Statistical Computation

One success of statistical science and, in particular, the analysis of environmental data is the richness of our models to describe spatial dependence, including the influence of covariates and the distribution of extreme observations. However, these more complex models, although attractive, have come with computational challenges that make them difficult to use. Data analysis with these extensions using traditional statistical computation is time consuming and difficult to implement. The computational technology that has been developed with AI applications in mind can be harnessed to advance statistical science.

Highlighted Presentations

TUESDAY, AUGUST 4

Reproducing Expert Judgement with Shortened Surveys

Patient-reported outcome measures allow us to measure constructs that are often not directly observable. Shortening questionnaires reduces respondent burden and cost and can increase data quality. However, shortened forms may have lower precision in latent trait estimation, but can they be used for accurate prediction of a diagnosis or expert judgment? We use a Markov Chain Monte Carlo algorithm to find shortened forms that maintain diagnostic accuracy.

Quantifying the Effect of Super Shoes on Elite Marathoning

Super shoes are lightweight running shoes designed with a carbon fiber plate for improving running economy. Since their widespread adoption by elite marathon runners in 2017, the men's and women's world records have each been broken three times. To quantify the impact of super shoes on professional marathon times, we apply a mixed effects model to 2011–2023 Chicago marathon data. Using our model, we compare the probability of top elite runners pre- and post-super shoes breaking the elusive 2:00 barrier (for males) and 2:10 barrier (for females).

Quantifying Changes in Extreme Snow Accumulations to Inform Infrastructure Design

Structures in snow-prone areas must be designed to withstand the weight of snow that accumulates on the roof. The design calculations require probabilistic estimates of annual extreme snow water equivalent, which are then used in a structural reliability analysis to determine the appropriate strength of the structural members (i.e., beams, columns, etc.) to prevent collapse during seasons of extreme snow accumulation. This work highlights the opportunities and challenges associated with blending data and expertise between the statistics, climate science, and structural engineering communities to improve United States' infrastructure design standards.

Highlighted Presentations

JSM2026

WEDNESDAY, AUGUST 5

Bridging Data Gaps in Oncology: Integrating Large Language Models and Recommendation Systems for Evidence Synthesis

The rapid growth of oncology research has generated an enormous volume of data. However, the information is often fragmented across datasets, registries, and publications, limiting its full utility in evidence synthesis and precision oncology. This work establishes a proof of concept demonstrating that the combination of LLMs with sophisticated recommendation algorithms can systematically identify novel and clinically plausible cancer treatments. This integrated approach may accelerate the identification of effective therapies for cancer patients, ultimately improving patient outcomes.

Digital Phenotyping for Mixed-Type mHealth Data

Modern mobile health assessment combines self-reported measures of participants' well-being with passively collected health behavior data from wearable devices throughout the day. Motivated by novel mixed-type functional data, we propose a multivariate functional principal component analysis approach using a semiparametric Gaussian copula model, assuming a multivariate generalized latent non-paranormal process as the underlying mechanism. The method is applied to ecological momentary assessment data from 307 participants in the National Institute of Mental Health Family Study of Mood and Affective Spectrum Disorders comprising their mood, anxiousness, energy, and activity. Our results uncover the dominant time-varying latent components across the four domains and recover clinically meaningful digital phenotypes for mood disorder stratification.

What the Body Whispers: Predicting Brain Health Trajectories from Everyday Smartwatch Data

Wearable technologies enable continuous, low-burden collection of high-frequency behavioral, physiological, and circadian signals in everyday life, creating new opportunities to study brain health outside clinical settings. However, predicting brain-health trajectories from such data remains challenging due to missingness, temporal dependence, and substantial heterogeneity across individuals. This work presents a framework for the prediction of cognitive and affective functioning from sleep, physical activity, heart-rate dynamics, and environmental exposures. The analysis centers on quantifying prediction error and model reliability under real-world conditions.

Highlighted Presentations

THURSDAY, AUGUST 6

Seeing Above and Below the Canopy: Modeling and Interpreting Species Occupancy with Deep Multi-Modal Habitat Representations

Effective conservation and restoration of species is an increasingly urgent priority. Here we show how limitations can be addressed by incorporating AI-derived, multi-modal habitat representations from both overhead satellite imagery and ground-level camera-trap imagery. Across diverse geography and species, these learned representations yield more accurate out-of-sample predictions than models based on conventional environmental covariates alone, and combining satellite and ground-level views provides complementary gains. Our approach provides a scalable path toward microhabitat-aware and interpretable species–habitat models that can better support restoration planning and management decisions. We provide code and model weights to reproduce our results.

Recent Advances in High-Performance Statistical Computing Algorithms

The majorization-minimization principle is an extremely general framework for deriving optimization algorithms. The MM principle finds wide applications in large-scale machine learning problems such as matrix completion, discriminant analysis, and nonnegative matrix factorizations. This work presents novel applications of the MM principle in the big data setting, including parallel block least squares, de-weighting weighted least squares, large-scale variance component model, independent component analysis, and multi-level Monte Carlo.

From Connected Vehicle Data to Safer Roads: Applications for State DOTs

Government agencies increasingly use connected vehicle and telematics data to monitor roadway conditions and improve safety, but transforming massive, complex datasets into actionable insights remains a challenge. This work presents a scalable data pipeline that converts terabytes of raw trip and event data into analysis-ready datasets using high-performance computing. The work highlights practical challenges in data integration, computational efficiency, and interpretation.