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.