Journal of Data Science 23 (3), 451–453
July 2025

DOI: 10.6339/25-JDS233EDI
Editorial

Editorial: 2024 WNAR/IMS/Graybill Annual Meeting

Tianjian Zhou[1,*], Brian Wiens[2], and Tianying Wang[1]

[1] *Department of Statistics, Colorado State University, USA*
[2] *Rivus Pharmaceuticals, USA*

The WNAR/IMS Annual Meeting is a longstanding conference jointly organized by the Western North American Region (WNAR) of the International Biometric Society and the Institute of Mathematical Statistics (IMS). Originally created to foster networking among researchers in the western regions of the United States and Canada, the meeting has since grown to attract participants from across North America and around the globe. It encompasses a broad spectrum of topics in statistics and data science, reflecting the diverse interests of its participants. The Graybill Conference was initiated in 2001 by the Department of Statistics at Colorado State University (CSU) to honor Professor Franklin A. Graybill, the department's founding chair. Each iteration of the Graybill Conference is centered around a specific theme, which varies from year to year to reflect pressing challenges and emerging areas in statistics. The 2024 theme was "Rare Disease Drug Development".

In 2024, the WNAR/IMS Annual Meeting and the Graybill Conference were held jointly on the CSU campus in Fort Collins, Colorado, USA. The event attracted a total of 435 participants, setting new attendance records for both conferences. This special issue features 7 peer-reviewed articles authored by participants of the joint conference. These contributions showcase recent advances in statistics and data science, reflecting the vitality and breadth of the discussions that took place during the event.

Data Science Review Wang et al. (2025b) review semiparametric methods used in case-control studies for gene-environment interactions. They focus on two predominant methodological frameworks: retrospective likelihood and hypothetical population, both under the assumption of gene-environment independence. By deriving a new estimator from both perspectives, the authors connect the two frameworks, shedding light on their fundamental similarities and differences. The proposed estimator is straightforward and simple to implement. A numerical study demonstrates its validity and empirical efficiency gains, opening the door to further theoretical investigations.

Statistical Data Science Jin and Leroux (2025) compare estimators of discriminative performance for time-to-event outcomes, in particular the time-dependent area under the receiver operating characteristic curve and concordance, in the context of the Cox proportional hazards model. The authors examine two classes of estimators: semi-parametric and non-parametric. They identify a previously unrecognized flaw in certain semi-parametric estimators, which can substantially overestimate out-of-sample performance due to a bias favoring overfitted models. On the other hand, non-parametric estimators do not exhibit this behavior but suffer from high variability. To mitigate the instability issue, the authors propose smoothing the estimates over time with penalized regression splines. Their findings are illustrated using data from the National Health and Nutrition Examination Survey.

[*] Corresponding author. Email: tianjian.zhou@colostate.edu.

Shan (2025) assesses methods for analyzing time-to-event endpoints when the proportional hazards assumption may not apply. When the proportional hazards model applies, analysis with the Cox model will be most efficient; when other configurations apply, other analyses may be preferred. The author compares analysis with the Cox model, the logrank text, the Wilcoxon test and restricted mean survival time for required sample size, power, and clinical interpretation, and make recommendations on the use of each.

Wang et al. (2025a) propose a knockoff-based variable selection approach tailored to address missing data and measurement error in high-dimensional metabolomics studies, providing guaranteed false discovery rate control. Extensive simulation studies confirm that the method maintains power across various scenarios of measurement error and missingness. Applying the approach to serum and urine metabolite profiles from the Women's Health Initiative, the authors identify a concise set of lipid metabolites consistently associated with breast and colorectal cancer risks, demonstrating the method's utility for robust biomarker discovery from noisy real-world data.

Chen et al. (2025) consider observational studies of pathways from exposure to outcome, when a very large number of baseline characteristics are collected and considered as potential confounders. The best strategy for reducing the baseline characteristic variables to a manageable number is not always clear. The authors consider and compare various selection strategies to evaluate estimation of target parameters. The methods are applied to an assessment of how the gut microbiome, which contains a huge number of marker genes, may affect cognitive function in elderly individuals.

Pollock et al. (2025) present a method for reducing bias from preferential sampling when pooled testing is used to assess presence of a marker. The motivating example is presence of disease in wild animal populations. While spatial sampling can provide a more random sample, and testing each individual animal can provide information on individual subjects, both are inefficient when the animals tend to cluster in large packs at a few places. The authors combine prior work on each of these issues into a novel method, applied to coronavirus infection among bats in California, and use a simulation study to demonstrate lack of bias.

Data Science in Action Ghosh et al. (2025) present a five-phase piecewise-linear change-point model designed to characterize longitudinal medical cost trajectories. Estimated using a grid search coupled with penalized generalized estimating equations, the model segments spending patterns into pre-disease, diagnostic, intensive-treatment, stable, and terminal phases. An analysis of pancreatic cancer patients from a large SEER–Medicare cohort identifies distinct cost patterns: a brief surge around diagnosis, an intensive initial treatment phase, a subsequent stable period, and a final escalation preceding death. Comorbidities and surgical interventions predominantly drive costs in earlier phases, while geographic location exerts only a minor influence. This framework provides policymakers with a clear and actionable method for forecasting and managing healthcare expenditures across disease phases.

We would like to express our sincere gratitude to the Editor, Jun Yan, for providing us with the opportunity to share the discussions from the 2024 WNAR/IMS/Graybill Annual Meeting with a broader audience. We are also deeply grateful to the anonymous reviewers for their timely and thoughtful evaluations of the submitted manuscripts. Lastly, we extend our appreciation to the conference committee, including Prince Allotey and Catherine Lee (WNAR Program Chairs), Jie Peng (IMS Program Chair), Jingling Ye (Graybill Program Chair), and Wen Zhou (Local Organizing Committee Chair), for their efforts in making the joint conference a success.

References

Chen M, Nguyen TT, Liu J (2025). High-dimensional confounding in causal mediation: A comparison study of double machine learning and regularized partial correlation network. *Journal of Data Science*, 23(3): 521–541. https://doi.org/10.6339/25-JDS1169

Ghosh I, Zheng Q, Egger M, Kong M (2025). Estimating healthcare expenditure using parametric change point models. *Journal of Data Science*, 23(3): 560–574. https://doi.org/10.6339/24-JDS1157

Jin Y, Leroux A (2025). Comparing estimators of discriminative performance of time-to-event models. *Journal of Data Science*, 23(3): 470–490. https://doi.org/10.6339/25-JDS1163

Pollock CP, Hoegh A, Irvine KM, de Wit LA, Reichert BE (2025). Estimating disease prevalence from preferentially sampled, pooled data. *Journal of Data Science*, 23(3): 542–559. https://doi.org/10.6339/25-JDS1191

Shan G (2025). Restricted mean survival time for a randomized study with survival outcome. *Journal of Data Science*, 23(3): 491–498. https://doi.org/10.6339/25-JDS1177

Wang R, Dai R, Huang Y, Neuhouser M, Lampe J, Raftery D, et al. (2025a). Variable selection with FDR control for noisy data–an application to screening metabolites that are associated with breast and colorectal cancer. *Journal of Data Science*, 23(3): 499–520. https://doi.org/10.6339/25-JDS1166

Wang T, Liu J, Wu A (2025b). Bibliographical connections for semiparametric analysis in case-control studies on gene-environment interactions. *Journal of Data Science*, 23(3): 454–469. https://doi.org/10.6339/24-JDS1155