The American Statistical Association (ASA) New Jersey Chapter and Bayer Statistics and Data Insights 7th Annual Workshop

The topic this year: ***Statistical Data Mining in the Pharmaceutical Industry.***

When: Friday November 8, 2019

Where: Bayer Pharmaceuticals, 100 Bayer Boulevard, Whippany NJ 07981

Tentative Schedule:

- **11AM-1PM:** poster session, lunch and welcome (lunch will be served from 11:30 AM - 12:30 PM)

- **1PM-5PM:** Keynote Presentation, Panel Discussion, Short Course, and 1 break

The event is free of charge. Onsite parking is available. If you plan on attending, you must RSVP to Nancy O'Donnell (email: nancy.odonnell@bayer.com) by **October 14, 2019.** Please note that the workshop is limited to 130 attendees due to the conference room capacity. We encourage early registration since once capacity is reached, registration will be closed.
Organizers are soliciting interested participants to present posters at this workshop. The poster topic is preferred but NOT limited to the workshop theme. Posters previously presented elsewhere are welcome as well. The poster session will be limited to 10 posters and will be held from 11:00 AM to 12:45 PM. If you are interested in submitting a poster for consideration, please send your poster title and abstract to Shivani Nanda (Shivani.nanda@bayer.com) by **October 1, 2019.** You will be notified of whether your poster has been accepted shortly thereafter.

Speakers and Presentations:

- **Keynote Speaker:** David Madigan, Professor of Statistics at Columbia University in New York City

Towards Honest Inference from Real-world Healthcare Data
(joint work with George Hripcsak, Patrick Ryan, Martijn Schuemie, and Marc Suchard)
In practice, our learning healthcare system relies primarily on observational studies generating one effect estimate at a time using customized study designs with unknown operating characteristics and publishing – or not – one estimate at a time. When we investigate the distribution of estimates that this process has produced, we see clear evidence of its shortcomings, including an apparent over-abundance of statistically significant effects. We propose a standardized process for performing observational research that can be evaluated, calibrated and applied at scale to generate a more reliable and complete evidence base than previously possible. We demonstrate this new paradigm by generating evidence about all pairwise comparisons of 39 treatments for hypertension for a relevant set of 58 health outcomes using nine large-scale health record databases from four countries. In total, we estimate 1.3M hazard ratios, each using a comparative effectiveness study design and propensity score stratification on par with current one-off observational studies in the literature. Moreover, the process enables us to employ negative and positive controls to evaluate and calibrate estimates ensuring, for example, that the 95% confidence interval includes the true effect size 95% of time. The result set consistently reflects current established knowledge where known, and its distribution shows no evidence of the faults of the current process.

- **Panel Discussion**: ***Applications of Data Mining Methods in Clinical Development***

Mercedeh Ghadessi (Bayer)
Vanja Vlajnic (Bayer)
David Madigan (Columbia University)
Ilya Lipkovich (Eli Lilly)
Bohdana Ratitch (Eli Lilly)
José Pinheiro (Janssen)
David James (Novartis)

- **Short Course**: Ilya Lipkovich and Bohdana Ratitch (Eli Lilly)

Statistical Data Mining of Clinical Data
Data mining/machine learning (DMML) methods are becoming an integral part of data analysis in many application domains, including clinical drug development. A wealth of data is being collected during a clinical development program but these data are often underutilized. We advocate for a principled use of machine learning (as opposed to haphazard data-dredging) to more fully utilize information from existing clinical trial data. In this short course we highlight how DMML approaches differ from classical statistics and give examples of successful applications of DMML methods in clinical research. We provide an overview of selected methods for supervised learning (prediction) and semi-supervised learning (subgroup identification). We illustrate the former methods with an application to predicting avoidable dropouts in clinical trials and the latter with simulated data on for a clinical trial that failed to show efficacy in the overall population. We discuss how DMML can be integrated in the overall clinical data analysis plan to ensure a principled analysis process.

Course Outline
- Data mining/machine learning and classical statistics
- Why data mining for clinical data? Typical applications
- A sample of methods for supervised learning with application to predicting preventable dropouts
- Overview of methods for subgroup identification from clinical data with application to a failed Phase 3 trial
- Integrating data mining plan in study development program
- Summary, Q & A.