Biostatistical Contributions to the Use of Machine Learning in Regulatory Science

Di Zhang[1], Jaejoon Song[1], Yong Ma[1], Sai Dharmarajan[1], Hana Lee[1], Rongmei Zhang[1], Tae Hyun Jung[1], Bingqi Han[2], Mark Levenson[1]

[1]Division of Biometrics VII, Office of Biostatistics, Center for Drug Evaluation and Research, FDA, [2]George Washington University

Introduction

Background

- In post-market drug safety surveillance, machine learning (ML) has been used to support regulatory decision making.
- ML was used in prediction and causal inference problems.

Overview of Methods

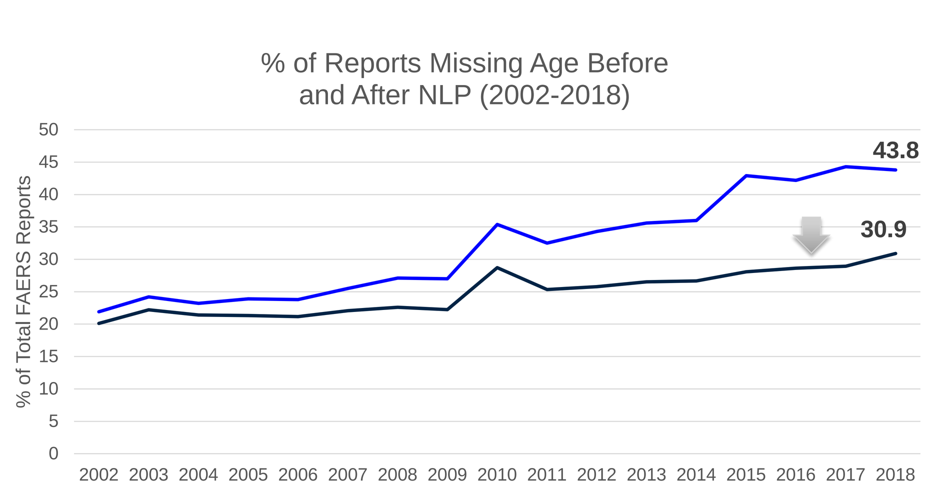
- **Supervised versus Unsupervised Learning**
  - Supervised: CART, penalized regression, Bayesian additive regression tree (BART), random forest (RF), boosting, support vector machine (SVM)
  - Unsupervised: clustering
- **Super Learner (SL)**
  - Utilize multiple ML algorithms
- **Natural Language Processing (NLP)**
  - Rule-based

Variable Ascertainment

GOAL: ascertain missing values for critical variables, such as age, in the FDA Adverse Event Reporting System (FAERS) database

METHOD: a rule-based NLP tool was developed to impute the missing age based on the unstructured free-text narratives

FINDING: high performance of this NLP tool with sensitivity of 0.99, specificity of 0.93, and PPV of 0.82



Health Outcome Identification (HOI)

GOAL: to improve algorithms to identify health outcomes in FDA Sentinel system

METHOD: applied various ML methods to classify anaphylaxis outcome based on claim-derived covariates, and EHR-derived covariates using NLP rule-based approaches

FINDING: combination use of Bayesian additive regression trees and NLP approaches improved the performance of the algorithm compared to NLP approaches only

CHALLENGE: limited data with known class labels available for training ML models, especially for rare safety outcomes

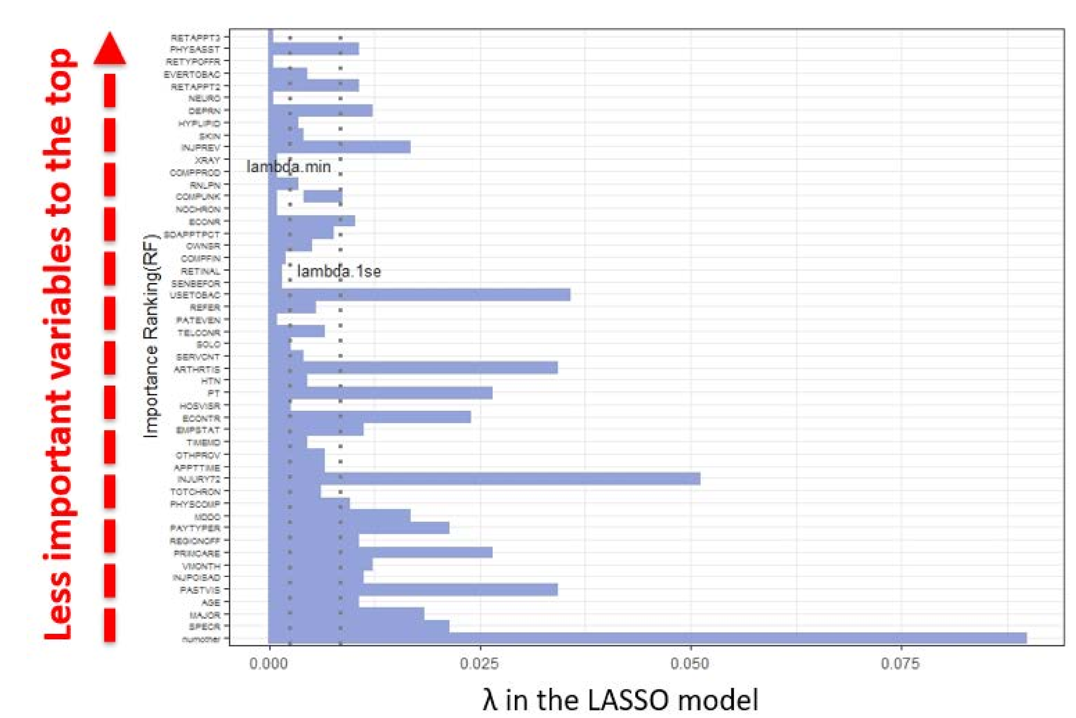
EXTENSION: determine the optimal training data size for imbalanced data applying ML methods for HOI, using a model-based method and a learning curve method

Risk Factor Identification

GOAL: to understand factors driving opioid prescribing using National Ambulatory Medical Care Survey (NAMCS) data, which aims to reduce preventable harm from potentially inappropriate opioid prescriptions

METHOD: LASSO penalized logistic regression, RF and SVM

FINDING: the identified risk factors (e.g., patients with arthritis were more likely prescribed with opioid) can inform inappropriate opioid prescribing



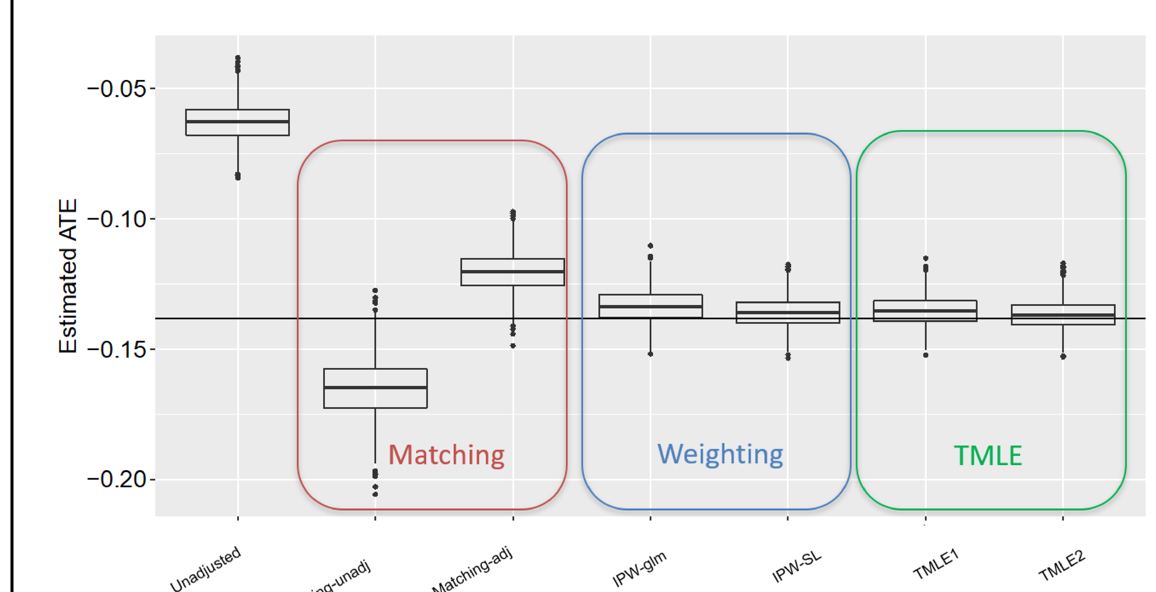
Causal Inference

GOAL: to evaluate the utility of targeted learning (TL) framework for establishing roadmap for optimally estimating causal effects and association measures from real world data (RWD)

METHOD: apply targeted minimum loss-based estimation (TMLE) combined with SL in both randomized trials and complex observational settings

FINDING:

- TMLE + SL outperformed the propensity score (PS) matching and inverse probability weighting methods using parametric modeling
- SL for estimating PS or missingness probability improved the overall performance of causal effect estimation in both PS matching and weighting analyses



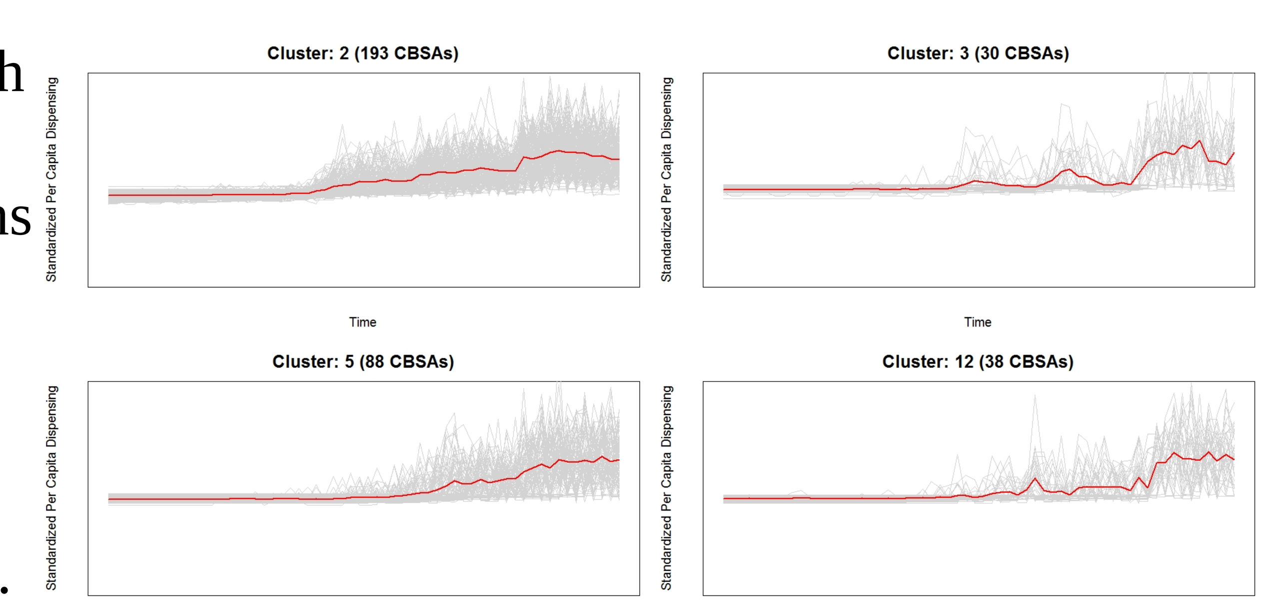
Drug Utilization Pattern Discovery

GOAL: to explore pharmacy dispensing data to provide insights for oversight of drug utilization

METHOD: developed tool, geoMapr, to analyze proprietary, nationally projected data for prescription drug dispensing and applied in PHAST PM database

- Decision tree-based ensemble algorithms to explore geographical factors associated with prescription dispensing
- Clustering for discovery of temporal patterns

FINDING: ML helps illustrate common trajectories of prescription change among geographic areas, and identify important factors contributed for the trajectory over time.



Conclusion

- Use of ML methods could improve performance of model predictions and causal effect estimation to better inform regulatory decision making.
- These projects demonstrate the present utility and future potential of ML for regulatory science.