Annotated Bibliography

1. Badurdeen, D. et al. (2012). “Timing of procedure and compliance with outpatient endoscopy among an underserved population in an inner-city tertiary institution”. Annals of Epidemiology, 22(7).

The study examined the impact of timing on keeping an appointment. The main predictors were season, day, and afternoon vs. morning appointments. The outcome is compliance. Compliant patients are those that showed up for their procedures vs. those not showing up without canceling or rescheduling their procedures up to 24 hours.

They controlled for age, sex, marital status, and insurance type. They did not report the control variable effects.

Sample: 2,873 patients who were scheduled for endoscopy (EGD, colonoscopy, or both) from September 2009 to August 2010.

The authors assessed compliance for all endoscopy procedures and colonoscopy-only procedures. They found seasonal differences in which warmer months had higher associations with compliance than the winter season for colonoscopy-only and “all-type” procedures. They found marginally significant differences for the time of day and colonoscopy-only encounters.

2. Boulestix, Anne-Laure, and Matthias Schmid. "Machine learning versus statistical modeling." *Biometrical Journal* 56.4 (2014): 588-93. Web.

This article provides a discussion of twin papers; they are: “Probability Estimation with machine learning methods for dichotomous and multicategory outcome: Theory” and “Probability Estimation with machine learning methods for dichotomous and multicategory outcome: Applications”. The purpose of the article is to compare the machine learning methods for probability estimation as presented and overviewed by the twin papers with statistical modeling, noting the weaknesses and issues of machine learning methods in biometrical practice; notably, choice of learning algorithm, parameter tuning, and computational transportability and reproducibility, are the issues that are addressed.

The purpose of the article is to provide understanding of the difference between explaining and predicting, that is, descriptive or predictive goals in developing models and rules in the biometrical field. While predictive purposes are well addressed with machine learning methods, interpretability becomes much more obscure and less important.

Each of the issues are addressed:

Choice of Learning Algorithm (“Degrees of Freedom” of the data analyst)

Because there are so many machine learning methods to choose from and it is agreed that no singular learning method (such as in statistical modeling where logistic modeling is primarily used) is the best over all possible datasets and predictive problems, the problem of choosing the best method yields two problems: over-optimism (“fishing for significance”) and parameter flexibility.

This first problem stems from incorrect cross-validation of possible methods and stumbling upon results yielding in high accuracy estimates. While there are correct cross-validation approaches and corrections, the sheer amount of methods to choose from can often lead to a biased selection from methods that have been cross-validated incorrectly.

The second problem of parameter flexibility is in part a look at the weakness of traditional statistical methods falling into model misspecification due to over-restrictive models, manually fit; the other side of the spectrum is that because parameters can be tuned in machine learning methods, the flexibility comes with computational cost to avoid model misspecification.

Computational transportability, reporting, and reproducibility

The application of models produced by machine learning methods to other researchers’ datasets and reports. Corresponding software objects and training data sets are elements necessary to be readily available or distributable when allowing others to use one’s machine learning model and adds computation cost and limitations to the receiver.

Reproducibility errors primarily stem from software differences in computation and lack of important and formal documentation of software code along with access to method implementation platforms.

3. Grömping, Ulrike. "Variable Importance Assessment in Regression: Linear Regression versus Random Forest." *The American Statistician* 63.4 (2009): 308-19. *JSTOR*. Web.

This article discusses variable importance and the difference in the two approaches that are used to evaluate variable importance: Linear Regression and Random Forest. The goal of the comparison is to help detect and be aware of different biases that may be introduced by either method. Much of the discussion compares various machine learning methods against one another.

The variable importance metrics used for linear regression are LMG (based off the average weighted distribution of R2 proposed by Lindeman, Merenda, and Gold) and PVMD (Proportional Marginal Variance Decomposition) and are compared with variable importance assessments in random forests.

The simulation study conducted compared how the increase in the mtry argument of the random forest function gave relative importance to variables to the LMG and PVMD methods of variable importance used in linear regression.

4. Mozaffari-Kermani, Mehran, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K. Jha. "Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare." *IEEE Journal of Biomedical and Health Informatics* 19.6 (2015): 1893-905. Web.

This article provides an in-depth analysis of a commonly implemented attack on machine learning methods as applied to healthcare data known as a poisoning attack. The authors provide evidence for the effectiveness of this type of attack and provides countermeasures to resolve poisoned data if security is already breached.

The motive behind this paper is to spur further research to defend against and counter poisoning attacks on machine learning.

The study done involved six machine learning methods as applied to five different data sets that are common in healthcare and are classification problems in and of themselves. The machine learning methods are: Best-First decision tree, Ripple-down rule learner, Naïve Bayes Decision Tree, Nearest-Neighbor Classifier, Multilayer Perceptron, and Sequential Minimal Optimization. The type of data sets used were as follows: Thyroid Disease, Breast Cancer, Acute Inflammations, Echocardiogram, and Molecular Biology.

The attack model utilizes two algorithms and is algorithm-independent so that simply switching machine learning algorithms is not a plausible countermeasure nor an effective one for that matter. The results indicate that adding up to 20% malicious instances each algorithm lost 15-20% classification accuracy. It poses 4 fronts on which healthcare data can be attacked: Attacking without Access to the Training Data, Attacking Unknown Machine-Learning Algorithms, Adapting the Attacks to Real Patterns, Attacking n-way Classification. Each of these fronts has particularly vulnerable or robust machine learning methods.

The countermeasure provides a way to identify suspicious malicious instances using one algorithm with consideration of also being able to identify non-malicious changes in classification accuracy.

5. Steiner, John F., Michael R. Shainline, Mary Christine Bishop, and Stan Xu. "Reducing Missed Primary Care Appointments in a Learning Health System." *Medical Care* 54.7 (2016): 689-96. Web.

This article discusses a study conducted with the operational goal of improving clinic efficiency and to provide predictive models to enhance future learning health system projections and disseminations. The predictors used were: age, sex, race, marital status, duration of KPO enrollment, employment status, health insurance type, lead time to appointment, day and time of appointment, distance from residence to clinic, comorbidity, having missed 1 or more appointments within 6 months of the index appointment, and an emergency department visit or hospitalization within 6 months of the index appointment.

The primary important predictor found was having missed 1 or more appointments within 6 months of the index appointment.

The primary outcome was whether the appointment was kept, missed, or cancelled.

The study was a replicated design having an index clinic (IC) and a replicated clinic (RC). The purpose was to compare the two clinics and to see if similarities in effectiveness of the intervention resulted and which would then be grounds to locally generalize and disseminate the system to all 25 of the healthcare clinics under the administration of Kaiser Permanente Colorado (KPCO).

There were observations with certain characteristics that were excluded from the collected data to be used for randomized intervention.

Sample size estimation was conducted to ensure the minimum sample size was attained based on the risk reduction parameter; after exclusions, 8804 and 7497 appointments were analyzed from IC and RC data respectively.

Randomized intervention was implemented as a treatment for each clinic individually. The intervention treatment was a IVR-T system (Intervention Voice Reminder or Text Message) system that reached a landline or cell phone via automated voice calls or text message, depending on the preference of the patient. The randomization algorithm came from the Structured Query Language and accounted for patients with multiple appointments (the first appointment became the index appointment for analysis though multiple appointments from patients may have be included in the randomized intervention).

The conclusion reached by the authors is that their intervention program significantly reduced the amount of missed appointments while increasing the amount of cancelled appointments for the group defined as high-risk (delineated by the top 25% of observed risk of missing appointments based on prior history of missing appointments). They found, however, that, for low-risk patients, though the intervention systems produced a positive reduction in the amount of appointments missed, this difference was not significant. The authors also discuss the importance to understand the localized context under which the study was conducted but also how their findings aligned with literature and other researches done before their study and outside of that same context. This external validation provides grounds for generalization and further research of other contributory variables and future implementations of intervention systems such as transportation assistance, repeated calls, and pre-appointment meetings, especially for high-risk patients. Randomization was concluded to be a source of credibility for the results concerning operational efficiency. The authors end by promoting “rapid, inexpensive, randomized interventions that rely exclusively on available data” for future health-care administrative systems.

6. Steyerberg, Ewout W., Tjeerd Van Der Ploeg, and Ben Van Calster. "Risk prediction with machine learning and regression methods." *Biometrical Journal*56.4 (2014): 601-06. Web.

This paper discusses twin papers from Kruppa et al. that look at machine learning methods both in theory and application. Steyerberg uses this paper to both summarize and analyze the two papers that heavily advocate machine learning implementation into biomedical research and study. The overall discussion provides a comparison with logistic regression modeling and the pros and cons of each with regards to medical data. The paper is split into 4 sections discussing current modern regression modeling techniques, model uncertainty and parsimony, prediction sensibility and interpretability, and future-looking discussion of machine learning techniques in coincidence with logistic regression, respectively.

The number of parameters in machine learning that can be tuned which allow for flexibility and in some ways complexity control. There is however, comparable techniques namely splines and shrinkage penalizations that perform very well if not better than tuned machine learning parameters. The computational cost for machine learning and sometimes lack of interpretability seems then to not be necessary.

The uncertainty of prediction models, though a concern for both machine learning techniques and regression modeling, remains primarily a problem for data hungry machine learning techniques, that is, nonparametric machine learning techniques that may require large datasets to perform well. However, when medical research datasets are reasonably sized and are not of high order non-linear complexity, logistic regression (with penalization) and machine learning techniques have been shown to perform similarly to one another.

The issue that remains is that many medical datasets are rather small which is a problem in and of itself especially awhen the number p predictors is greater than the number n observations. For both logistic regression and machine learning, this poses a problem in overfitting in some cases by selecting too many important predictors or biasing coefficients because near significant predictors were not valued as important.

The concluding table compares 3 machine learning techniques and logistic regression based on consistency, flexibility, sample size requirements, and interpretability. Nearest Neighbor (NN) has good consistency and flexibility with heavy sample size requirements and lack of interpretability. Random Forest (RF) mostly has good consistency (especially when applied to a linear regression problem), good flexibility with the number of parameters available to tune, decent sample size requirement, and poor interpretability. Support Vector Machines (SVM) share the same traits as RF. Logistic Regression (Logreg) has poor consistency, decent flexibility with cubic splines (rcs) or fractional polynomials (FP) and shrinkage penalizations (lasso, ridge), very low sample size requirement, and great interpretability.

In conclusion, the author maintains that logistic regression will remain the default modeling approach to probability estimation in medical risk prediction, though machine learning techniques may be useful for highly complex order problems, though they are infrequent in medical data, and can still be used as a comparison method to regression results.