Overview of Applications and Added Value of Statistical Learning and Artificial Intelligence (AI) in Drug Development

Richard Baumgartner

2018 ASA Biopharmaceutical Section Regulatory -Industry Statistics Workshop

Acknowledgements

- Amanda Zhao, Robin Mogg (BARDS, Early Clinical Statistics, Merck)
- Devan Mehrotra (Biostatistics and Research Decision Sciences, BARDS, Early Development Statistics, Merck)
- Yue-Ming Chen, Vladimir Svetnik, Dan Holder (BARDS, Biometrics Research, Merck)

- Jinghua He, Mehmet Burcu, Sarah Liu, Ed Bortnichak (Pharmacoepidemiology, Center for Observational Studies and Real World Evidence (CORE), Merck)
- Weifeng Xu (Statistical Programming, BARDS, Merck)

Outline

- Introduction and application of statistical learning in drug development
- Weak rare model and signal screening
- Ensemble learning and medical claims phenotyping
- Emerging applications
  - Subgroup analyses with incorporation of machine learning methods
  - Conformal predictors
  - Deep learning using high-dimensional data such as medical images across multiple applications for segmentation and prediction
- Conclusions

Introduction
Some Concepts Germane to Statistical Learning

- Machine Learning –
  - *Constructs algorithms that can learn from data.*
- Statistical Learning –
  - *Is a branch of applied statistics that emerged in response to machine learning.*
  - *Emphasizing statistical models and assessment of uncertainty*
- Data Science –
  - *Is the extraction of knowledge from data using ideas from mathematics, statistics, machine learning, computer science engineering, and ….*
- All of these are very similar ….with different emphases

(from Trevor Hastie, 2015)

Statistical Learning in Pharmaceutical Industry is Being Applied across All Stages of Drug Development

- Drug Discovery:
  - Prediction of Compound Activity in quantitative structure-activity relationship (QSAR)
- Preclinical Development:
  - Segmentation in imaging assays for applications in preclinical efficacy and safety
- Prediction Challenges in Clinical Development:
  - Personalized (precision) medicine (responder vs. non-responder analysis)
  - Optimal treatment regime recommendations and subgroup analysis
  - Optimization of clinical trial execution
  - Clinical Safety and Risk Monitoring
- Real World Evidence and Observational Studies:
  - Phenotyping medical claims data
  - Heterogeneous treatment estimation in causal inference
  - Personalized healthcare
  - Applications in digital health using sensor or streaming data
  - Pharmacovigilance
- …

Feature Screening in a Rare Weak Model

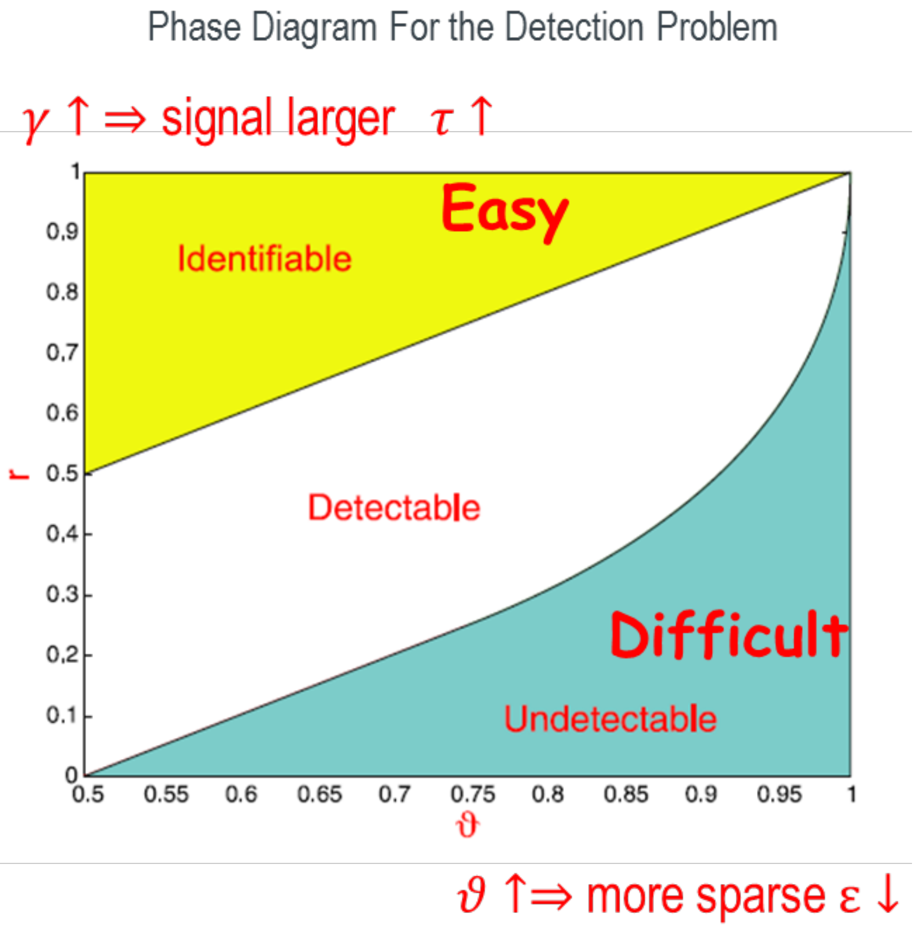
- Challenge: How to detect weak and rare signals in multidimensional biomedical data (feature identification and selection for predictive modeling) given that large effect sizes are rare in biology

- “There can be *some* large and predictable effects on behavior, but not a lot, because, if there were, then these different effects would interfere with each other, and as a result it would be hard to see any consistent effects of anything in observational data. The analogy is to a fish tank full of piranhas: it won’t take long before they eat each other.”

Andrew Gelman

http://andrewgelman.com/2017/12/15/piranha-problem-social-psychology-behavioral-economics-button-pushing-model-science-eats/

Rare and Weak Model and Interplay between Sparsity and Effect Magnitude for Informative Feature Detection



γ ↑ ⇒ signal larger τ ↑

ϑ ↑⇒ more sparse ε ↓

- To conduct an overall test of complete null hypothesis, testing whether *all* test statistics are distributed N (0, 1):

$$H_0{}^{(m)}\colon X_i\ i.i.d. \sim N\ (0,1), 1 \le i \le m$$

Against an alternative that a small fraction is distributed as normal with a nonzero mean τ:

$$H_1{}^{(m)}\colon X_i\ i.i.d.\ \sim (1-\varepsilon) N\ (0,1) \\ +\ \varepsilon N(\tau, 1), \\ 1 \le i \le m$$

- Two key parameters:
  - ε the fraction of the non-null effects/ sparsity;
  - τ the nonzero effect sizes/ signal strength

- Watershed effect: Sparsity vs Signal Strength plot contains distinct partitions with different properties in terms of feature detection

False Discovery Rate and Higher Criticism Thresholding

- Benjamini-Hochberg FDR:

– For p values $P^m = (P_1, P_2, \dots, P_m)$ for the m tests. Let $P_{(0)} < P_{(1)} < \cdots < P_{(m)}$. The BH threshold is defined for pre-specified $0 < \alpha < 1$ as

$$T_{BH} = \max\left\{P_{(i)}\colon\ P_{(i)} \leq \alpha \frac{i}{m}, 0 \leq i \leq m\right\}.$$

- Local FDR:

– Define the two component mixture model in terms of the density of the individual p values as $\mathrm{f}(x) = \eta_0 f_0(x) + (1 - \eta_0) f_A(x)$; Using Bayes' rule, the local FDR

– $$fdr(x) = Pr(\text{"null"}|X = x) = \frac{\eta_0 f_0(x)}{f(x)} = \frac{\eta_0}{f(x)}.$$

- Higher Criticism (HC)

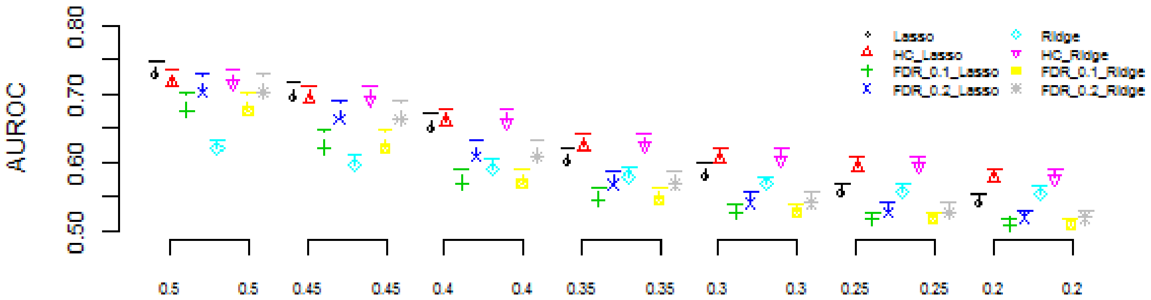
– Arranging the p-values from the smallest to largest $p_{(1)}, \dots, p_{(m)}$, define the higher criticism objective function

- $$\widehat{HC}(p_{(i)}) = \frac{|\hat{F}(x) - x|}{\sqrt{\hat{F}(x)(1 - \hat{F}(x))/m}} = \frac{|\frac{i}{m} - p_{(i)}|}{\sqrt{\frac{i}{m} * (1 - \frac{i}{m})/m}}.$$

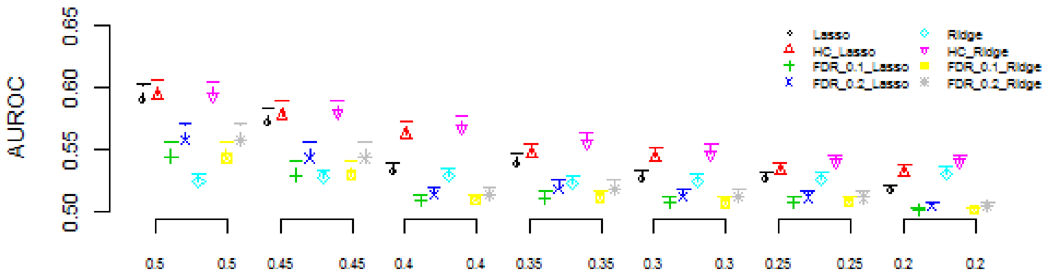
– The maximum of the HC objective function is obtained and the corresponding p value is taken as the HC decision threshold for signal detection.

Feature Selection and Classification

Scenario a: $\mathrm{N} = 100, p = 500, \mathrm{nz} = 10, \varepsilon = 0.02$.



Scenario b: $\mathrm{N} = 100, p = 10000, \mathrm{nz} = 10, \varepsilon = 0.001$.



- HC threshold in combination with Lasso or Ridge has an overall better performance compared to Lasso or Ridge alone or FDR + Lasso/ Ridge. While raising FDR cutoff helps the performance when signals are stronger, the improvement of applying HC threshold over FDR threshold is impressive when the dimension of features is larger and the signals are weaker.
- In the rare and weak settings, we need to select features in a way so that FDR is high, so that we are able to include more useful features for classification, which implies the potential application in biomarker screening and discovery.

Learn and Confirm Paradigm for Feature Screening and Validation

Learn (Discovery Data Set)		Confirm (Validation Data Set)	
Input	An initial high-dimensional panel of features (e.g. genes, spectral peaks, etc.)	Input	Lower dimensional signature obtained from the discovery data set
Feature Screening	Marginal testing of the features (genes) to obtain set of features that are associated with the outcome. HC cutoff applied (screening threshold)	Feature Validation	Signature obtained from the discovery data set will be profiled and validated FDR cutoff applied
Signature Construction	Statistical (machine) learning is used to evaluate the found set of features		

Challenges and Recommendations

- Challenges: Detection of rare/weak signals appears to be a ubiquitous across many disciplines. It poses challenges in feature identification, interpretation, and clinical utility.

- General recommendations: During the discovery phase with a large set of features to be screened and little knowledge of disease/ biological/ target-related mechanisms, applying HC method along with the commonly used BH-FDR will help capture weaker effects that could be of potential utility with further validation.

- Higher Criticism based methods are finding applications in the areas of supervised feature screening, unsupervised feature screening in clustering and Principal Component Analysis and anomaly detection

Phenotyping of the Medical Claims Data
Background / Motivation

- Administrative data is increasingly used for measurement of quality of care and outcomes by payers and healthcare organizations as a part of real world evidence (RWE) generation
- Examples of phenotyping, disease case ascertainment comprise of applications in oncology, heart failure, frailty, osteoporosis, etc.
- Cancer stage is the most important risk factor associated with survival and its ascertainment from the medical claims data is desirable to support RWE
  - Cancers are historically diagnosed at late stage
- Identification of cancer stage from administrative data a major challenge
  - Traditionally used algorithms based on decision trees deliver poorly on achieving high sensitivity/specificity simultaneously
- Machine Learning ensembles have been showing great promise to improve on the prediction/classification performance in the medical claims phenotyping due to large sample sizes (big data) that are conducive to their superior performance

Stacked Generalization
Superlearner

- Level-zero data: The original training set (X)
- Level-one data: The cross-validated predicted values (Z)
- Learner Library: Set of basis learners (learning algorithms)
- Meta-learner: Algorithm trained using
  - cross-validated predicted values Z and
  - the original target Y
  - typically linear

MERCK

Dense Random Effects Model for Classification Simulation I

- Dense Feature Assumption: Each predictor (dense) has a small independent random effect on the outcome (random)
- The expected signal strength is $E(||\delta^2||)=\alpha^2$
- Model Assumptions:

A. High Dimensional Asymptotics:

- The data X in R^{nxp} is generated as $X=Z\Sigma^{1/2}$
  - Entries of nxp matrix Z are i.i.d. with $E(Z_{ij})=0$, $Var(Z_{ij})=1$
  - Σ is a pxp deterministic matrix
- The sample size n → ∞ while the dimensionality p → ∞ as well, such that the aspect ratio p/n → $\gamma>0$

B. Random Weights for Classification

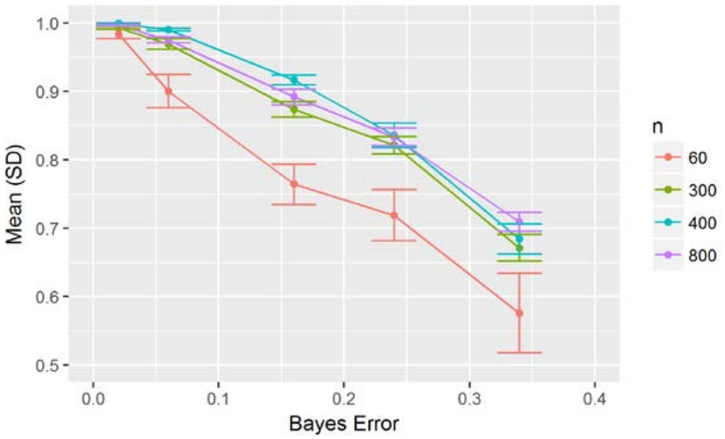
- Class centers (2 classes=-1/+1) μ_{-1} and μ_{+1} are randomly generated as $\mu_{-1}= \mu-\delta$ and $\mu_{-1}= \mu+\delta$, where
  - $E(\delta_i)=0$ and $Var(\delta_i)=\alpha^2/p$

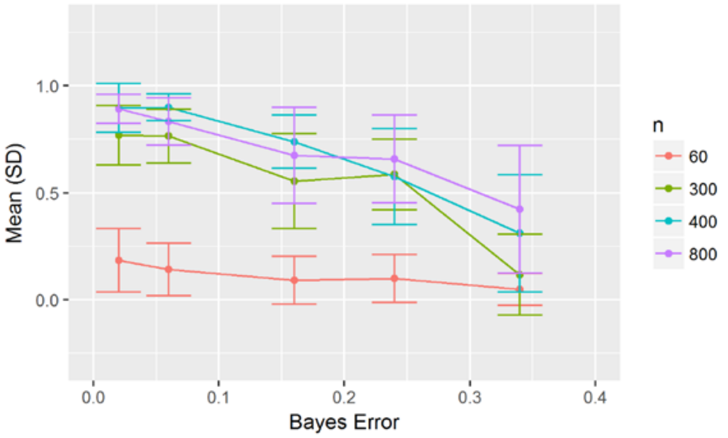
Simulation Study I: Dense 2 Class Linear Normal Model

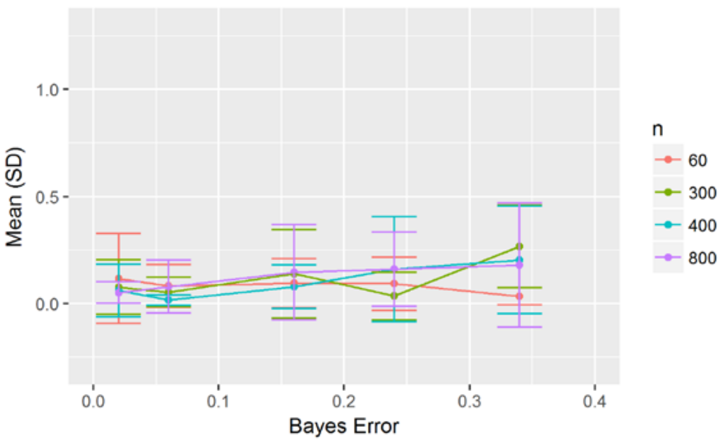
- Simulation setup:
  - p=40, AR(1), ρ=0.1
  - Ntrain=60, 300, 400, 800
  - Ntest=1000

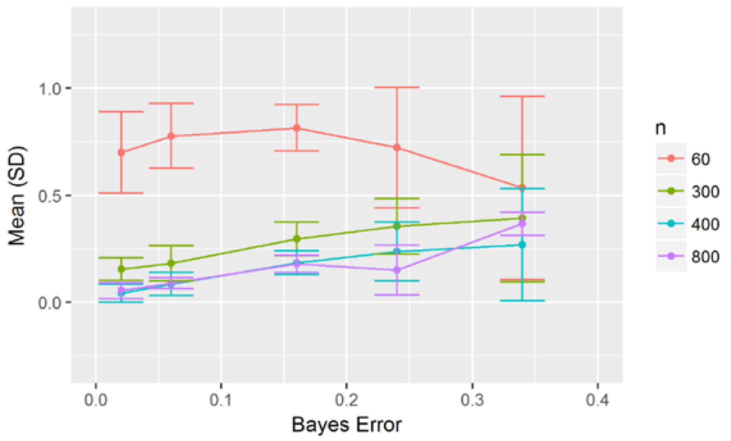
- Superlearner library:
  - LDA –linear discriminant analysis
  - LR – logistic regression
  - RF – random forest

Results: Simulation I


SL: AUROC


Coefficient:LDA


Coefficient:LR


Coefficient:RF

SL Coefficients provide insight into relative merit of component classifiers

Interesting interplay between Random Forest and LDA

MERCK

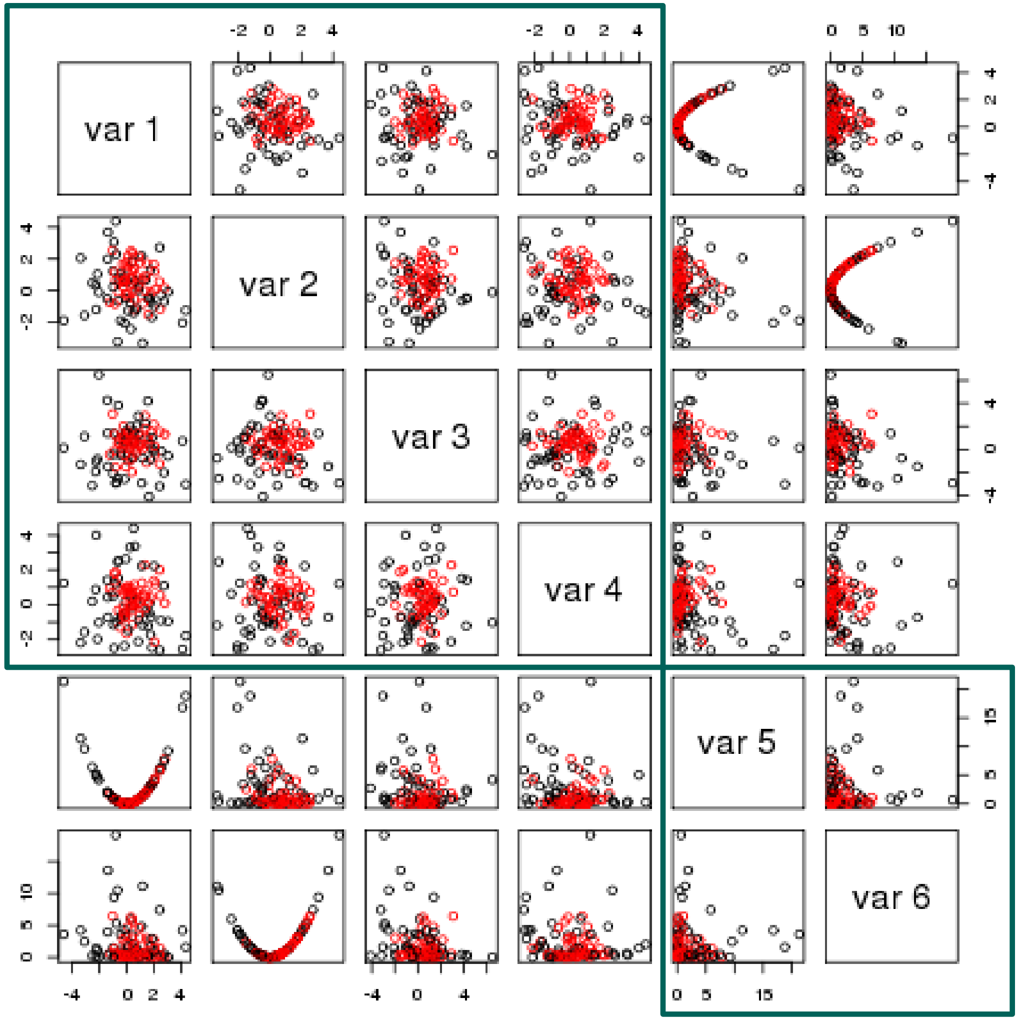
Simulation Study II:
Ringnorm Data Set (mlbench)

- Model: 2 Gaussian Distributions:
  - Class 1 multivariate normal with mean 0 and covariance 4 times the identity matrix
  - Class 2 has unit covariance and mean *(a,a,…,a)*

$$a = p^{-0.5}$$

- Simulation setup:
  - p=10
  - Number of Linearized Features: 0,2,4,…,10
  - Ntrain=60, 300, 400, 800
- Superlearner library:
  - LDA – linear discriminant analysis
  - LR – logistic regression
  - RF – random forest

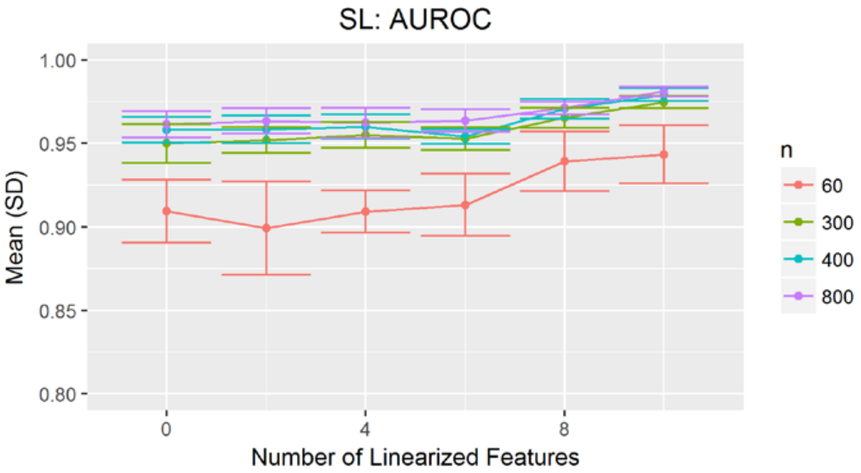
Data Plot

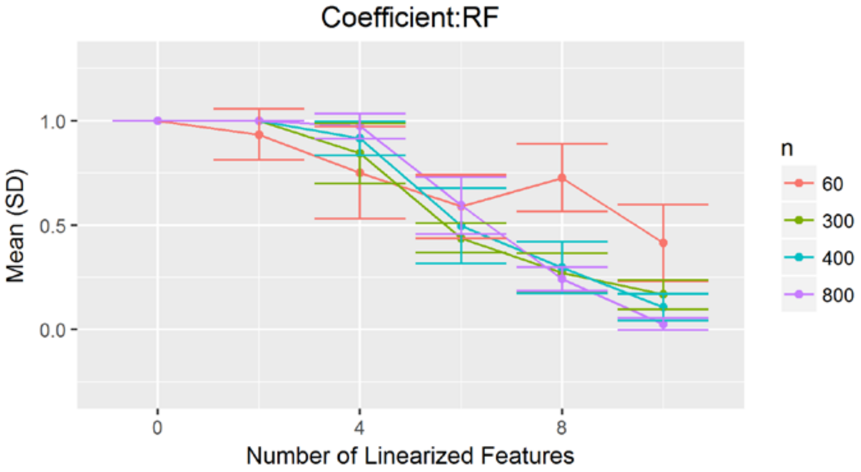


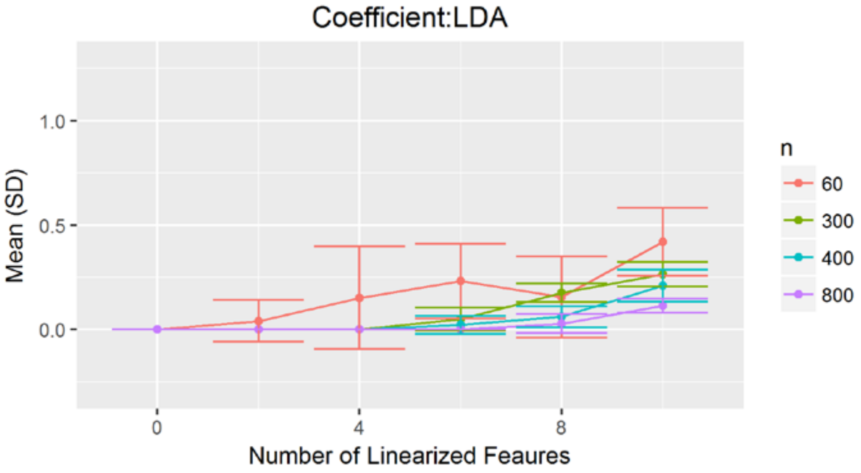
var1-var4: original first five variables

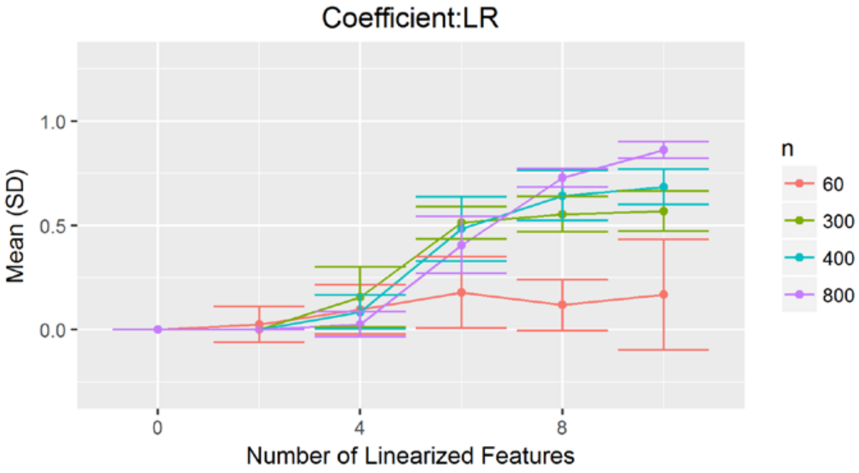
var5,var6: linearized variables

Results: Simulation II









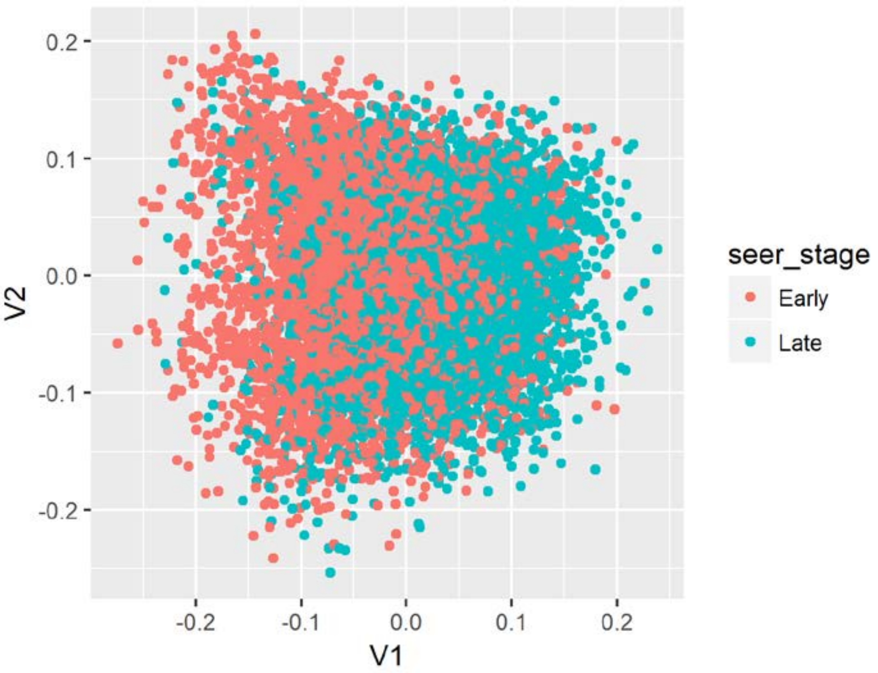
SEER and Medicare linked data

- A linkage between SEER cancer registry and Medicare administrative claims database
  - SEER is a large cancer registry that covers 25% of the US population
  - Medicare claims cover comprehensive inpatient/outpatient diagnosis and procedures received by Medicare beneficiaries
- Study included patients who were diagnosed with lung cancer and received chemotherapy in 2010-2011
  - Cancer stage classification algorithms were developed from Medicare inpatient/outpatient claims
  - Cancer stage classification from SEER registry served as gold standard for validation
    - Early: AJCC stage I/II/II (local)
    - Late: AJCC stage IV (metastatic)

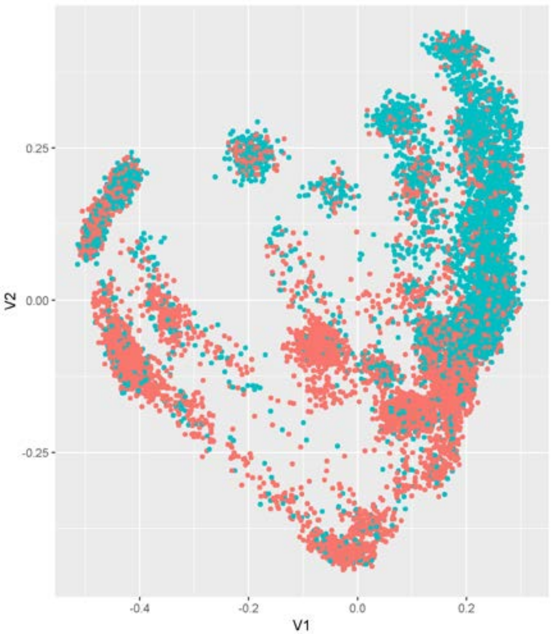
Study Cohort and Data Set Construction

- A total of 11,198 patients were included
- The constructed data set included three sets of predictors
    - C1: clinical variables (suggested by a clinical tree algorithm)
    - C2: demographics, lung surgery, radiation therapy, chemotherapy regimen, comorbidities
    - C3: lung and secondary malignancies diagnosis
- A total of 101 predictors
    - Qualitative (categorical): 68
    - Quantitative (continuous): 33
- Target class – late stage lung cancer according to SEER (gold standard)
    - Early: AJCC stage I/II/III n=6,039
    - Late: AJCC stage IV n=5,159

Results and Visualization by Unsupervised Random Forest



All variables projected to two dimensions denoted as V1 and V2



Top 5 variables obtained from random forest

Superlearner comprising of logistic Regression, xgboost and Random Forest achieved balanced sensitivity/specificity ~0.8

- C3_11_198_per: % metastases codes other site
- C24_3_noncranial_cnt: Number of claims for non-brain radiation
- C23_3_lobec: C23_3_lobec
- C3_8_met_cnt: Number of claims for metastases claims
- C3_9_196_per: % metastases codes lymph node site

Conclusions

- From the simulations
  - Ensembling by a superlearner shows good performance for larger data sets
  - Coefficients obtained from the superlearner are informative and provide insights about the data geometry
- From the real-world study
  - Current ML approaches significantly outperformed secondary cancer diagnosis
  - Random forest and superlearner exhibited superior performance in terms of sensitivity and specificity with respect to the logistic regression and xgboost
  - Superlearner also provided balanced sensitivity and specificity

Subgroup Analysis

- Recently there has been an active research in development of rigorous methods for finding subgroups of populations that are benefiting the treatment in both randomized clinical trials and observational studies. These include:
  - Decision trees/forest based methods
  - Bayesian methods
  - Methods for individual treatment regimen recommendation
- Pocock et al. 2002 argue that subgroup analysis procedure should begin with test for treatment-covariate interaction, as such test directly examines the strength of evidence for heterogeneity in treatment effect
- However, many studies are not sufficiently powered to identify a significant interaction as sufficient evidence that none exist
- Tree-based methods – naturally partition the input space, however there is potential overfitting
- Desiderata: simultaneous inferences regarding subpopulations
  - Statements that all members of the subpopulation satisfy, e.g. every member of a specific subpopulation benefits from the treatment

Example of Benefit - Safety Trade-off in an AD Trial Obtained from Bayesian Analysis Schnell et al. 2017

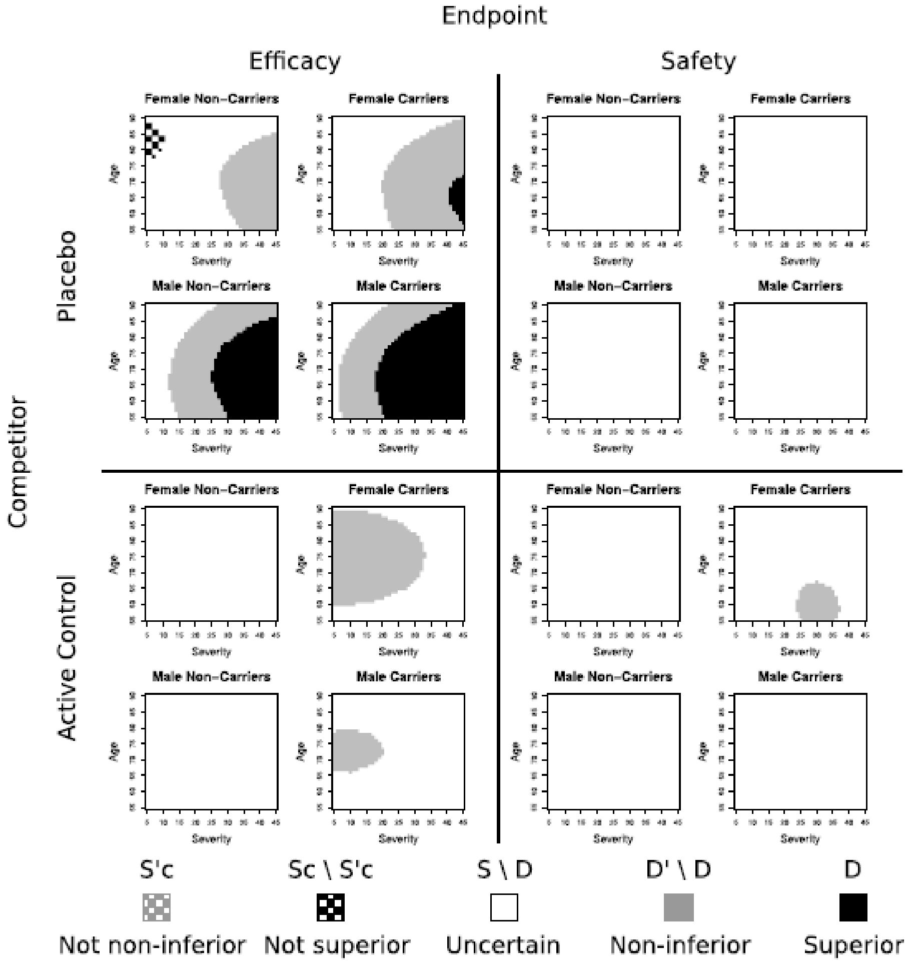


FIG. 2. *Individual 50% credible subgroup pairs plotted over the covariate space. Each large quadrant contains a plot of the credible subgroup pairs for that endpoint-competitor combination, with subgroups for noninferiority* (D′, S′) *and superiority* (D, S) *overlaid.*

Upper left: males with high disease severity tend to benefit from treatment vs. placebo

Bottom left: more non-inferiority in female and low severity patients vs. active control

Right hand side: uncertainty in the relative safety profiles, female carriers (of genetic biomarker) more promising for inferiority to active control

Active control and test treatment may both favor male and high-severity patients, potentially due to more activity of a similar mechanism

Conformal Inference

- Challenge: How to provide non-parametric prediction sets for binomial and continuous prediction outcomes
- Conformal inference:
  - Roots in Computer Science (proposed by Glenn Shafer and Vladimir Vovk) and currently further popularized by Larry Wasserman and Ryan Tibshirani in statistics
  - Based on probability theory and statistics using statistical tools such as p-values and frequentist concepts
  - Provides well calibrated prediction sets for individual predictions
  - Hallmarks of conformalization:
    - Given a training set, adding a sample in and obtaining a corresponding p-value
    - Test inversion to obtain the prediction interval
- Conformalization of usual algorithms:
  - Support vector machines
  - Random forests
  - Linear regression
  - Discriminant analysis
- Currently being applied in QSAR applications, with a potential utility in other areas of drug development

Leveraging Deep Neural Networks for Predictive Modeling using Multidimensional Imaging Inputs

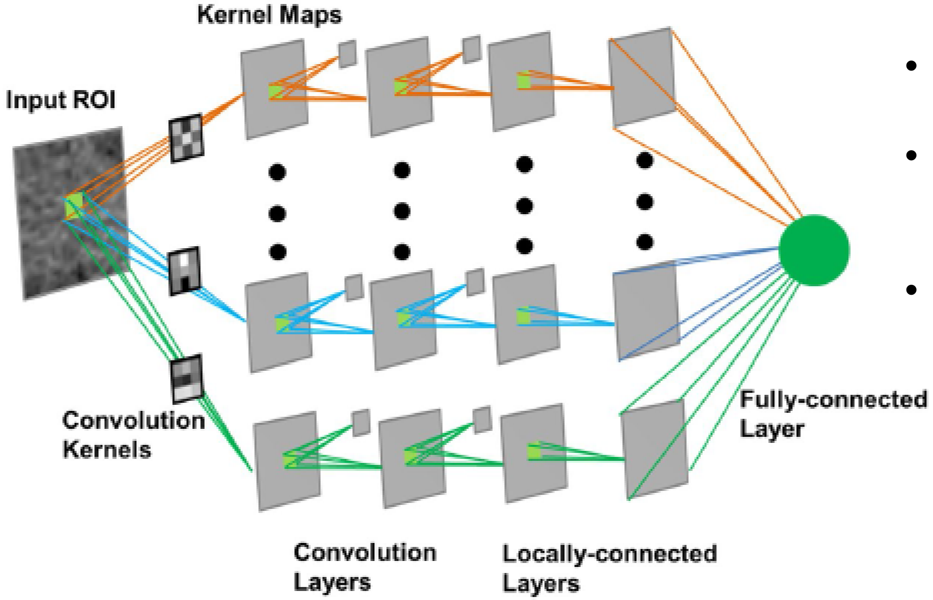


Figure 4. DL-CNN Structure. An input ROI is convolved with multiple convolution kernels, and the resulting values are collected into the corresponding kernel maps. This process repeats for several layers, giving the "deep" convolutional neural network. The network used in this study contains two convolution layers and two locally-connected layers, each of which contains 16 kernels.

- Example of bladder cancer treatment (chemotherapy) response prediction
- Extracted CT imaging regions of interest (ROIs) were directly used as inputs for deep learning without explicit feature engineering
- AUROC ~ 0.7

Cha et al. Scientific Reports, 2017

Final Considerations

- Machine learning advances are being reflected on and leveraged in statistical learning
- Applications of statistical learning span all stages of drug development
- Statistical learning provides added value in:
  - understanding and characterizing underlying data generating mechanisms
  - theoretical underpinnings and well defined probabilistic frameworks to facilitate development of interpretable statistical learning models
  - addressing potential biases ensuing in real world application of statistical learning methods
  - rigorous assessment of uncertainty to facilitate decision making
    - which in turn should translate into efficient treatment development and delivery to the patients
- Caveats: more sophisticated and complex methods need to be applied as fit for purpose and used in cases where they provide additional benefits to traditional analysis

References

- SL Bergquist et al. Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data. Proceedings of Machine Learning Research, 68:25-38, 2017

- Dobriban E and Wager S, High dimensional asymptotics in prediciton: Ridge regression and classification. Annals of Statistics 2018

- Lix L et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. Journal of Clinical Epidemiology, 61(12):1250-60, 2008.

- R. Baumgartner et al. Lung cancer stage ascertainment from the population-based administrative databases. PhilaSUG Meeting, Philadelphia, PA, 2018.

- Donoho D and Jin J. Higher criticism thresholding: optimal feature selection when useful features are rare and weak. PNAS 105(39): 14790-5, 2008

- Shafer G and Vovk V. A tutorial on conformal prediction. Journal of Machine Learning Research, 371-421, 2008.

- Lei J et al. Distribution-free predictive inference for regression. Journal of the American Statistical Association, in press, 2018.

- Schell P et al. Subgroup inference for multiple treatments and multiple endpoints in an Alzheimer's disease treatment trial. The Annals of Applied Statistics, 11(2), 949-66, 2017.

- Pocock SJ et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting. Statistics in medicine, 21(19), 2917-30, 2002.

- Fu H et al. Estimating optimal treatment regimes via subgroup identification in clinical trials and observational studies. Statistics in Medicine, 35(19), 3285-3302, 2016

- Wang Y et al. JASA, Learning Optimal Personalized Treatment Rules in Consideration of Benefit and Risk: With an Application to Treating Type 2 Diabetes Patients With Insulin Therapies. JASA, 113(521), 1-13,2018

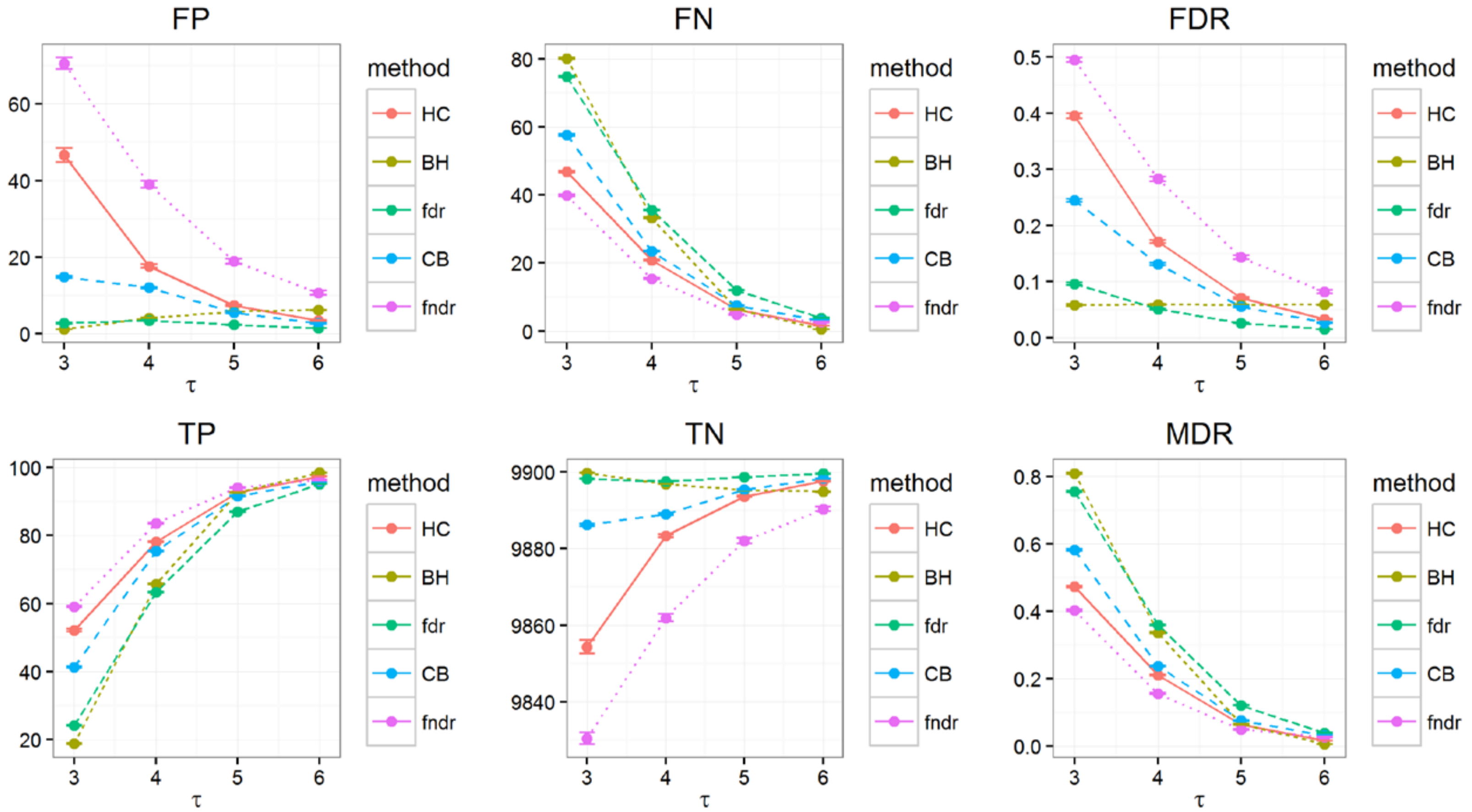
- Cha et al. Bladder Cancer treatment response assessment in CT using radiomics with deep learning, Scientific Reports, 7:8738, 2017.

Backups

MERCK

Feature Screening

Scenario: $\tau \in [3, 6]$ and $\varepsilon = 0.01$ on the signal identification boundary.



- HC has a higher false feature discovery rate but a low feature missed detection rate when the signals are more sparse and weaker.
- As the signals become easier to detect, HC performs similar to FDR controlled method (CB or local fdr cutoff = 0.5).
- BH-FDR ensures FDR to be well-controlled regardless of the signal strength, but for screening it is missing most of signals when the signals are rare and weak.

Conformal Inference: Toy Example (Czuber's problem) (Shafer & Vovk 2007)

- Consider 19 integers: 17, 20,10,17, 12, 15,19,22,17,19,14,22,18,17,13,12,18,15,17
  - $\min(Y_i) = 10, \max(Y_i) = 22$
- Goal: Find a prediction (confidence) set for the 20th number to be observed (n=19)

1. Create an augmented set by adding in a hypothetical y: 17, 20,10,17, 12, 15, 19, 22, 17, 19, 14, 22, 18, 17, 13, 12, 18, 15, 17, y

   $\bar{Y}_y = \frac{1}{n+1}(\sum_{i=1}^{n} Y_i + y)$=$\frac{1}{20}(314 + \mathrm{y})$ - obtain an average **including** y

2. Non-conformity score (residual) for y: $R_{n+1} = \left|\bar{Y}_y - y\right| = \frac{1}{20}|314 - 19y|$

3. Non-conformity score (residual) for Y_i: $R_i = \left|\bar{Y}_y - Y_i\right|$=$\frac{1}{20}|314 + y - 20Y_i|$

Under the Null hypothesis that $Y_{n+1} = y$, the 20 observations are exchangeable and each of them is equally likely as the other to be largest (i.e. the ranks of the residuals follow discrete uniform distribution)

- Since there are 20 numbers, there is a 19/20 (95%) chance that the R_{n+1} will not exceed the largest of the R_i -s
- We can write: $R_{n+1} \leq \max\{R_{Y_{max}}, R_{Y_{min}}\}$ or
  - $\frac{1}{20}|314 - 19y| \leq \max\{\frac{1}{20}|314+y-20*22|, \frac{1}{20}|314+y-20*10|\}$
- Therefore: $10 \leq \mathrm{y} \leq 24$ and the 95% prediction set for the sample will be [10,24]
- This interval is essentially the same to Fisher's interval (Shafer & Vovk 2007)
- The realization of Y_{20} turned out to be 16 falling between 10 and 24