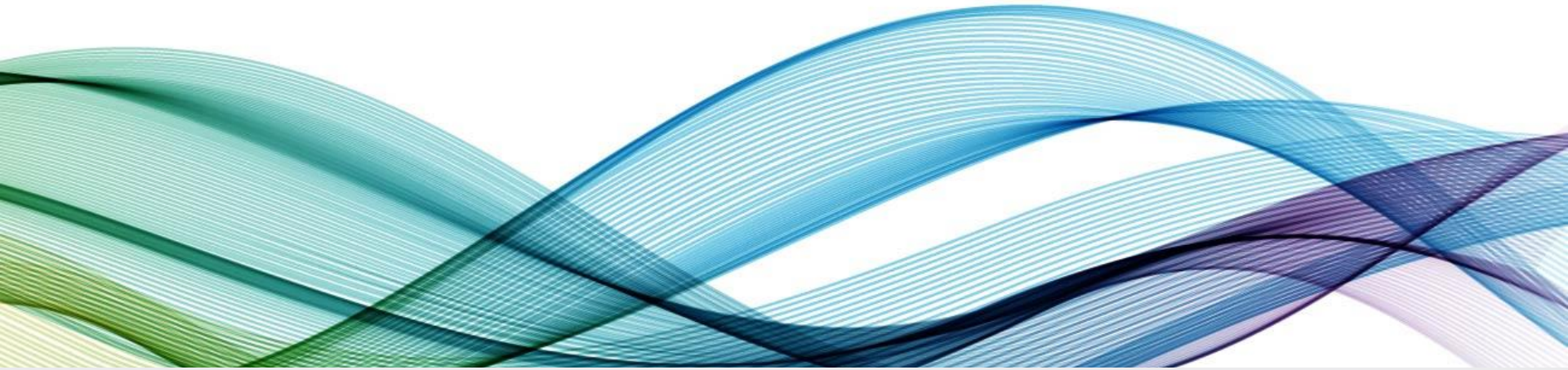
Prevention of Missing Data in Clinical Trials Informed by Statistical Learning from Past Clinical Trials

Bohdana Ratitch, QuintilesIMS
Ilya Lipkovich, QuintilesIMS

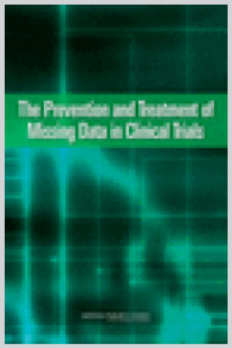
24th Annual Biopharmaceutical Applied Statistics Symposium
October 23, 2017



Outline

- Looking back at the NRC 2010 report
- Motivating example – case study in developing targeted retention strategies for HIV trials
- Identifying important predictors of potentially avoidable dropout
  - Traditional statistical methods
  - Machine (statistical) learning methods: basic concepts and some methods
- Variable Importance measures
- Interactions
- Partial Dependence Plots
- Summary, discussion

Prevention of Drop-out and Loss to Follow-up



The Prevention and Treatment of Missing Data in Clinical Trials

"… Two approaches to the problem are to **reduce the frequency of missing data** in the first place and to use appropriate statistical techniques that account for the missing data. The former approach is **preferred**,…

Prevention Theme in the NRC 2010 Report

- NRC 2010 report contains 4 recommendations related to minimizing missing data, e.g.,

  – Recommendation 2: Investigators, sponsors, and regulators should design clinical trials consistent with the goal of maximizing the number of participants who are maintained on the protocol-specified intervention until the outcome data are collected.

  – Recommendation 6: Study sponsors should explicitly anticipate potential problems of missing data. In particular, the trial protocol should contain a section that addresses missing data issues, including the anticipated amount of missing data, and steps taken in trial design and trial conduct to monitor and limit the impact of missing data.

Some Prevention Strategies in Trial Design and Conduct

Foster adherence to treatment and continuing study participation

without compromising safety and wellbeing of patients

- Minimize participants' burden and inconvenience
- Flexible consenting, e.g., with an option for reduced assessment burden
- Use technology to facilitate visit scheduling/reminders
- Offer an extension study with experimental drug
- Offer treatments to manage non-critical side effects
- Flexible dosing, if possible
- Consider "enriched" design (e.g., with a run-in period) or randomized withdrawal design
- Work with patients and investigators to convey importance of full participation
- Do not encourage discontinuation in case of protocol violations
- Outline effective communication and follow-up strategies in the protocol
- Use historic retention rates (in addition to historic enrollment rates) for site selection
- Retention strategies, including targeted strategies for subjects at high risk of loss to FU

Prevention Theme in the NRC 2010 Report (Cont'd)

- On-trial elicitation of intention to continue
- " …investigators and site personnel can collect information on which participants are at risk for dropping out and why: formal "intent-to-attend" questioning may help to identify reasons for dropout (see, e.g., Leon et al., 2007) and may yield useful covariates in missing data models. Factors influencing decisions to participate include:
  - (i) time and duration of visits,
  - (ii) need for assistance with transportation or child care,
  - (iii) need for reminders,
  - (iv) problems in relations with the staff,
  - (v) problems with blood drawing or other procedures,
  - (vi) side effects, and
  - (vii) perceptions of intervention efficacy."

Prevention Theme in the NRC 2010 Report (Cont'd)

- NRC 2010 contains a section

"**UNDERSTANDING THE CAUSES AND DEGREE OF DROPOUTS IN CLINICAL TRIALS**"

> "A crucial issue that sponsors must wrestle with in planning a clinical trial is … how much could be reduced through the use of various techniques … and consequently if they implement these various techniques, what degree of missingness is likely to remain.
>
> … Information from previously collected clinical studies would help in answering these questions."

Missing Data - Not Just a Statistician's Problem

LaVange & Permutt (2015):

"*We also believed that too much was being asked of statistical analysis. Trialists were sometimes complacent about missing data, assuming that the problem would be satisfactorily addressed in analysis. We hoped that a frank discussion by expert statisticians of the limitations of statistics would encourage improvements in the design and conduct of trials to lessen the need for analytical solutions based upon unproven assumptions.*"

Prevention of Missing Data - Is it a statistician's job too?

Seems like ClinOps issues, why are we talking about it here?

Proactive Strategies for Patient Retention

MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1528

Published online 17 July 2012 in Wiley Online Library

The statistician's role in the prevention of missing data

Sara Hughes,[a*] Julia Harris,[b] Nancy Flack,[c] and Robert L. Cuffe[d]

- A successful case study by GSK to develop a proactive plan for focused retention efforts in HIV trials
- From previously completed studies, identify:
  - **Treatment-related / unavoidable treatment discontinuations,** e.g., adverse events, protocol-defined insufficient efficacy
  - **Potentially avoidable drop-out,** e.g., loss to follow-up, subject decision, non-compliance with protocol, etc.
- Identify demographic characteristics that are strongly correlated with potentially avoidable drop-out
- Develop patient retention strategies focused on patients at high risk of potentially avoidable drop-out

GSK Case Study – Retrospective Analysis

- Pooled analysis of 4 studies in HIV previously completed by GSK
  - Overall 24% discontinuation rate, ~17% potentially avoidable drop-out
- Identified several significant predictors of potentially avoidable drop-out
  - Ethnicity: Non-white at higher risk than white
  - HIV transmission mode: IV drug use and heterosexual at higher risk vs others
  - Age: Younger subjects (< 38 years) at higher risk

Table III. Translating population results into individual risk.

One-year risk of non-treatment related discontinuation					
		Not intravenous drug use		Intravenous drug use	
Non-white	Age (yrs)	Homosexual	Heterosexual	Homosexual	Heterosexual
	≤38	18%	24%	21%	28%
	39-44	12%	16%	15%	20%
	>44	9%	12%	11%	14%
White	≤31	10%	13%	24%	31%
	32-44	8%	10%	19%	25%
	>44	7%	10%	18%	24%

Key:	≤10%	11-15%	16-20%	≥21%

GSK Case Study – Prevention in Future Trials

- Pooled analysis of 4 studies in HIV previously completed by GSK
  - Overall 24% discontinuation rate, ~17% potentially avoidable drop-out
- **Identified several significant predictors of potentially avoidable drop-out**
  - Ethnicity: Non-white at higher risk than white
  - HIV transmission mode: IV drug use and heterosexual at higher risk vs others
  - Age: Younger subjects (< 38 years) at higher risk
- **Action with future studies**
  - "Typical profile" of a future trial subject at higher risk of dropout
  - Development of country-specific and site-specific plans targeted to high-risk subjects
- **Effect in future studies**
  - **In a new Phase IIb trial: 2% non-treatment related dropout after 48 weeks (vs ~17% historical)**
  - Evaluation is ongoing in a Phase III program…

GSK Case Study - Potential Impact of Increased Retention

- Increased efficiency / power / confidence in trial conclusions
- With consistent reductions of drop-out, can potentially design future trials with smaller sample sizes, e.g.,

Table IV. Illustrating the potential impact of increased retention in HIV non-inferiority clinical trials.

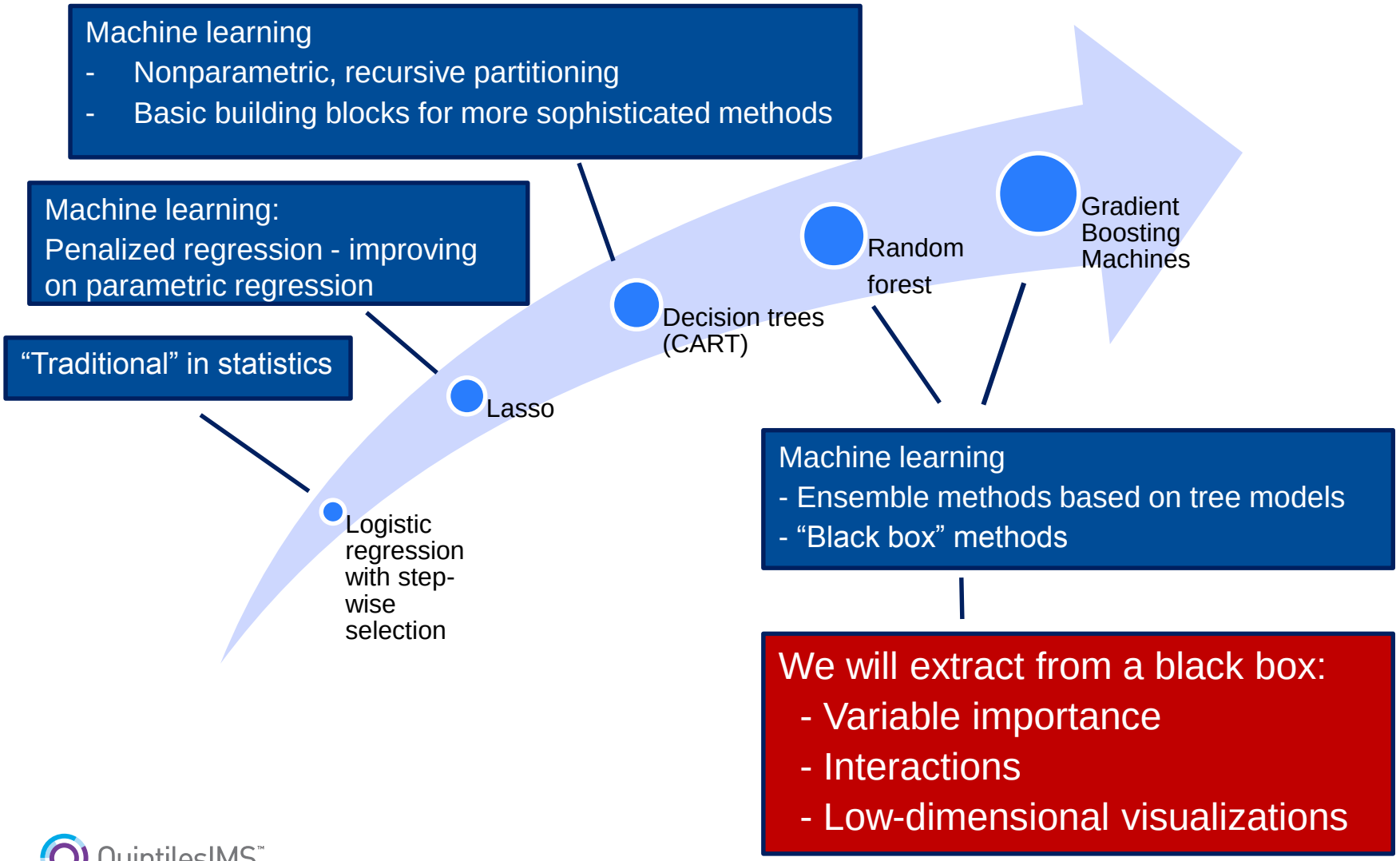
Increased retention levels	Control arm response rate†	Power (for fixed $N = 395$ per group)‡	N per group (for 90% power)‡
Base case	75% §	90%	395
5%	80%	94%	337 (15% reduction)
10%	85%	98%	268 (32% reduction)

Base case: no prevented dropout, habitual dropout rate.

Objectives for This Presentation (…the rest of it)

- Explore and apply advanced methods for identification of demographic and baseline characteristics that are strongly correlated with potentially avoidable drop-out
- Our focus is on variable selection to assess which predictors are important and to glean the general shape of their relationship with the probability of dropping out
  - Our goal is not to build a prediction model that will be used in the future to predict the odds of potentially avoidable dropout on individual future patients
  - Rather, we want to understand the key elements of the mechanism behind such dropouts, so as to be able to prevent/reduce them in the future
- We will consider potential predictors individually and their 2-way interactions, so that relatively simple, interpretable subject profiles can be formulated for targeted retention strategies

Methods Considered in This Presentation



Example Dataset

- Patterned after typical studies in Type II diabetes
- 600 subjects in all treatment groups combined
  - 510 (85%) – completers
  - 90 (15%) - potentially avoidable dropout

  Note: unavoidable dropout is not included in this analysis
- 14 candidate baseline predictors:
  - AGE, SEX, RACE, ETHNIC
  - BLBMI, H_OBESITY – baseline BMI and history of obesity
  - BLSYSBP, BLDIABP – baseline blood pressure
  - H_NEUROPATHY, C_NEPHROPATHY – history of neuropathy and comorbid nephropathy
  - BLHBA1C, BLFPG – baseline HbA1c and fasting plasma glucose
  - BLEGFR – baseline EGFR (>=90 – normal)
  - YSD5CAT - <5 years vs >=5 years since diagnosis
- Interactions may be important, but don't know which in advance:
  - 91 potential 2-way interaction terms (given 2 categories of RACE)
- Total: 105 potential predictors

“Traditional” Statistical Variable Selection Approaches

- In the GSK study:
  - Cox proportional hazards model for time to potentially avoidable dropout was used to identify significant predictors among 6 candidates
  - One-year predicted risks of potentially avoidable dropout were calculated from the model for each combination of significant factors.
- Logistic regression to model a binary variable for potentially avoidable dropout using stepwise variable selection, e.g., based on Akaike’s information criterion (AIC) or Bayesian Information Criterion (BIC)
- (Generalized) linear models are sometimes thought to be a preferred method because of their interpretability

Logistic Regression with Stepwise Variable Selection

- Some of the stepwise selection drawbacks:
  - subjective choices for the selection/elimination significance levels
  - model uncertainty - multiple different parameter combinations may have a similar fit to the data
  - a greedy method - picks the variable which is the current "winner" in explaining residual variance
  - predictors can appear significant or not, depending on what other predictors are in the model
  - instability – small changes in the data may lead to drastic changes in results
  - dangers of multiple testing (p-values do not have usual interpretation as they don’t account for selection process)

Logistic Regression with Stepwise Variable Selection

- Some of the stepwise selection drawbacks:
  - subjective choices for the selection/elimination significance levels
  - model uncertainty - multiple different parameter combinations may have a similar fit to the data
  - a greedy method - picks the variable which is the current "winner" in explaining residual variance
  - predictors can appear significant or not, depending on what other predictors are in the model
  - instability – small changes in the data may lead to drastic changes in results
  - dangers of multiple testing (p-values do not have usual interpretation as they don't account for selection process)

- AIC and BIC are some of the criteria that can be used for selection decisions at each step
  - AIC=-2 logL + 2p
  - BIC=-2 logL + p*log(n)

  where p is the number of parameters in the model (penalty for increased model complexity, discourages overfitting) and n (in BIC) is the sample size

  - provide the means to compare models in terms of their fit to the data, favoring the models with fewer parameters
  - removes the need for the user to specify significance levels (although AIC is equivalent to ~15.7%)

Results from Stepwise Selection on Example Dataset

- Logistic regression using stepwise (both directions) selection with AIC

R package "MASS"

```
logistic.base<-glm(y.train~.,family=binomial(link='logit'),data=temp.train.scaled)
logistic.step<-stepAIC(logistic.base, direction="both")
```

- Number of variables selected: 39

```
y.train ~ SEX + ETHNIC + AGE + YSD5CAT + BLDIABP + BLFPG + BLHBA1C + BLEGFR +
`SEX:AGE` + `SEX:YSD5CAT` + `SEX:BLSYSBP` + `SEX:BLFPG` +
`ETHNIC:AGE` + `ETHNIC:BLBMI` + `ETHNIC:BLDIABP` + `ETHNIC:BLFPG` +
  `ETHNIC:H_NEUROPATHY` + `ETHNIC:RACE.WH`+
`AGE:BLSYSBP` + `AGE:BLFPG` + `AGE:BLHBA1C` + `AGE:BLEGFR` +
`BLBMI:YSD5CAT` +`BLBMI:BLSYSBP` + `BLBMI:BLDIABP` + `BLBMI:H_NEUROPATHY` +
`YSD5CAT:BLSYSBP` + `YSD5CAT:BLDIABP` + `YSD5CAT:BLEGFR` +
`BLSYSBP:BLDIABP` + `BLSYSBP:BLEGFR` + `BLSYSBP:H_OBESITY` +
`BLDIABP:BLFPG` + `BLDIABP:BLHBA1C` + `BLDIABP:BLEGFR` + `BLDIABP:H_NEUROPATHY` +
`BLHBA1C:H_NEUROPATHY` +
`BLEGFR:H_NEUROPATHY`
```

Results from Stepwise Selection on Example Dataset

- Logistic regression using stepwise (both directions) selection with BIC

R package "MASS"

```
logistic.base<-glm(y.train~.,family=binomial(link='logit'),data=temp.train.scaled)
logistic.step<-stepAIC(logistic.base, direction="both", k=log(nrow(temp.train.scaled))
```

- Number of variables selected: 39

```
y.train ~ SEX + ETHNIC + AGE + YSD5CAT + BLDIABP + BLFPG + BLHBA1C + BLEGFR +
`SEX:AGE` + `SEX:YSD5CAT` + `SEX:BLSYSBP` + `SEX:BLFPG` +
`ETHNIC:AGE` + `ETHNIC:BLBMI` + `ETHNIC:BLDIABP` + `ETHNIC:BLFPG` +
  `ETHNIC:H_NEUROPATHY` + `ETHNIC:RACE.WH`+
`AGE:BLSYSBP` + `AGE:BLFPG` + `AGE:BLHBA1C` + `AGE:BLEGFR` +
`BLBMI:YSD5CAT` +`BLBMI:BLSYSBP` + `BLBMI:BLDIABP` + `BLBMI:H_NEUROPATHY` +
`YSD5CAT:BLSYSBP` + `YSD5CAT:BLDIABP` + `YSD5CAT:BLEGFR` +
`BLSYSBP:BLDIABP` + `BLSYSBP:BLEGFR` + `BLSYSBP:H_OBESITY` +
`BLDIABP:BLFPG` + `BLDIABP:BLHBA1C` + `BLDIABP:BLEGFR` + `BLDIABP:H_NEUROPATHY` +
`BLHBA1C:H_NEUROPATHY` +
`BLEGFR:H_NEUROPATHY`
```

Results from Stepwise Selection

- With a large number of possibly correlated predictors,
  - a regression model can indicate how well the combination of predictors predicts the outcome variable, BUT
  - it may not give valid results about any individual predictor, about which predictors are redundant with respect to others, or their relative importance

- To design a targeted retention strategy, we need to narrow down the set of important predictors

Machine (Statistical) Learning (ML) Approaches

- What's attractive about them?
  - Good performance in the presence of a large number of candidate predictors
  - Parametric or non-parametric, flexible, some can "build in" interactions naturally
- Reservations against ML
  - Some approaches are thought of as "black boxes" – lack of interpretability
  - "Black art" to some statisticians
- So what?
  - We will try to dissect the black boxes: variable importance, explore interactions, low-dimensional visualizations
  - Easily available in R, some in SAS

Some Statistical Learning Jargon

- Learn a model = estimate a model = identify a model (element of discovery)
- Learning algorithm = method used to estimate parameters of a model
- Training set = dataset used for model estimation
- Test set = data used to evaluate predictive accuracy of an estimated model, typically not included in the training set
- Tuning parameters = meta-parameters that determine some aspects of the learning algorithm which the user needs to specify (strategies exist to identify/learn optimal settings for the data at hand)
- K-fold cross-validation: technique to evaluate predictive accuracy of a model or to select settings of tuning parameters:
  - Divide (partition) the training set into K folds (sets), typically randomly
  - Run learning K times, each time using all data except the k^{th} fold for training, and perform model evaluation on the k^{th} fold as the test set.
  - Average performance measure over K test sets.
  - For parameter tuning, perform the above steps with different parameter settings and select the one(s) providing the best average performance over K test sets.

Variable Importance

- Variable importance (VI) is a measure of the relative importance or contribution of a variable to predicting the response
- VI is used in many machine learning approaches where a single variable may contribute multiple times in different parts of the model – it provides a single score presenting its overall importance
- VI captures both the main effect of the variable and its involvement in interaction effects with other variables
- VI is defined in different ways that reflect the construction of specific types of learners … more in a bit

Parametric Machine Learning Methods

- Penalized regression is a parametric method that was developed to address some difficulties with the use of traditional regression methods in statistics in the presence of many predictors
- It can deal with a large number of noise covariates among candidate predictors minimizing their impact on estimation
- It is a less greedy approach compared to stepwise selection:
  - Loosely speaking, the idea is to "restrict" the best predictor by allowing it to explain only a portion of residual variance, i.e.
  - The predictor is not included in the model with its "full" LS coefficient but with the coefficient shrunk down to 0, giving more chance to other predictors later to kick in and explain what is left of the variance.

Penalized Regression

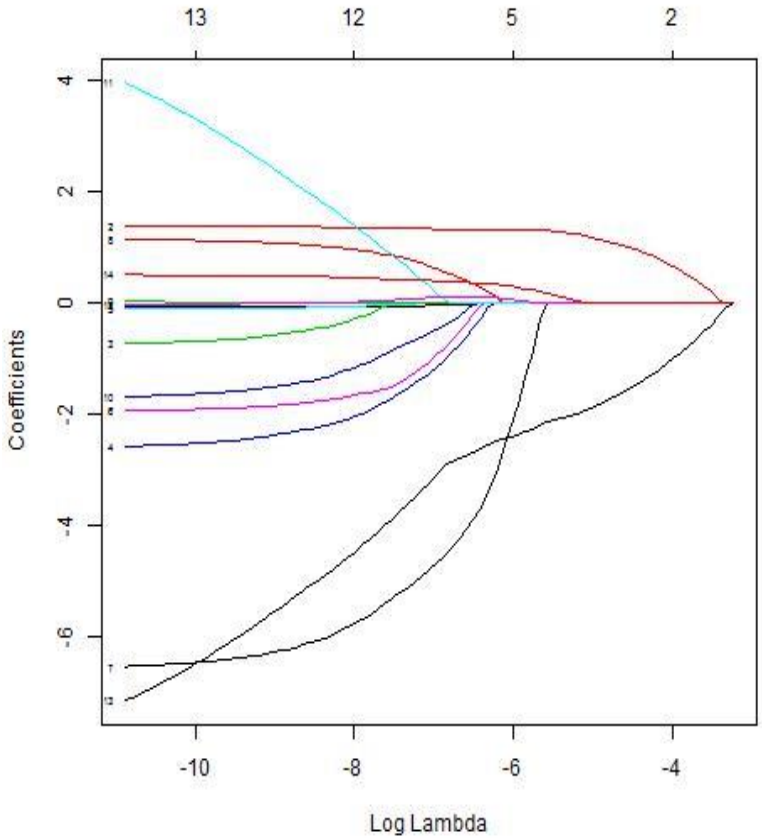
- Error (loss) function is minimized under a constraint that penalizes for model complexity and/or large absolute values of coefficients

$$\widetilde{\boldsymbol{\beta}} = arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{N}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right), \quad \text{subject to } Penalty(\boldsymbol{\beta}) < k.$$

  - Methods differ in terms of the penalty $Penalty(\boldsymbol{\beta})$ that they impose:
    - ridge regression (Hoerl and Kennard, 1970),
    - lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006)
    - elastic net (Zou and Hastie, 2005)
  - The two last methods would shrink estimated model coefficients either exactly to zero (effectively eliminating the variable from the model) or to some non-zero values.
  - The penalty causes parameter estimates to be biased but also decreases their variance – thus, variance-bias tradeoff.
  - Amount of shrinkage is determined by tuning parameters. Need to employ a tuning method to choose the optimal setting of the these parameters, e.g.,
    - cross-validation, or
    - AIC/BIC applied to the sequence of estimated models under varying tuning parameter

Results from Lasso on Example Dataset

- Penalty (lambda) versus coefficients plot when fitting main effects only



		Coefficients as lambda decreases					
(Intercept)	0	-1.7346	-1.69555	-1.69124	-2.02151	-1.97356	-1.24498
SEX	1	0	0	0	0	0	0
ETHNIC	2	0	0	0.07366	1.219443	1.291074	1.294679
AGE	3	0	0	0	0	0	0
BLBMI	4	0	0	0	0	0	0
YSD5CAT	5	0	0	0	0	0	0
BLSYSBP	6	0	0	0	0	0	0
BLDIABP	7	0	0	0	0	-0.30587	-1.36651
BLFPG	8	0	0	0	0	0	0
BLHBA1C	9	0	0	0	0	0	0
BLEGFR	10	0	0	0	0	0	0
C_NEPHROPATHY	11	0	0	0	0	0	0
H_NEUROPATHY	12	0	0	0	0	0	0.015474
H_OBESITY	13	0	-0.12786	-0.26599	-1.99742	-2.21154	-2.32229
RACE.WH	14	0	0	0	0.026115	0.197378	0.245308

Results from Lasso on Example Dataset

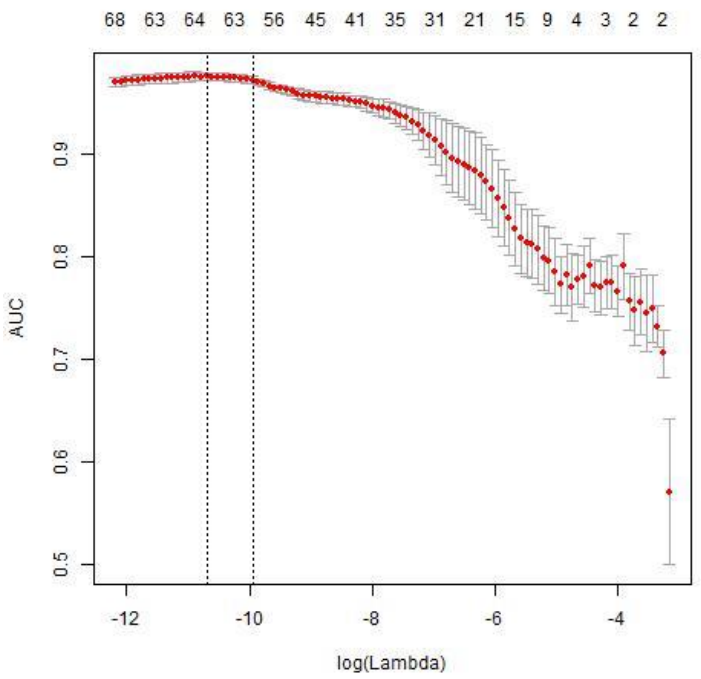
- Lasso with cross-validation to select the penalty parameter minimizing the cross-validation error

R package "glmnet"

```
lasso.cvfit<-cv.glmnet(x=as(pred.matrix.train.num.mat,"dgCMatrix"), y=y.train,
        standardize=FALSE, family = "binomial", type.measure="class", nfolds=5, alpha=1)
```

- Number of variables selected, on average, when minimizing misclassification error: 64
- Number of variables selected, on average, when maximizing AUC: 64
- Order in which variables are retained (with non-zero coefficients) as the penalty increases:

H_OBESITY
ETHNIC:H_NEUROPATHY
ETHNIC:RACE.WH
H_NEUROPATHY:RACE.WH
SEX:YSD5CAT
ETHNIC:AGE
SEX:ETHNIC
ETHNIC:YSD5CAT
BLDIABP:H_NEUROPATHY
SEX:C_NEPHROPATHY
YSD5CAT:C_NEPHROPATHY
SEX:H_NEUROPATHY
BLSYSBP:BLDIABP
SEX:BLEGFR
ETHNIC:BLBMI
BLSYSBP:RACE.WH
AGE:RACE.WH
...

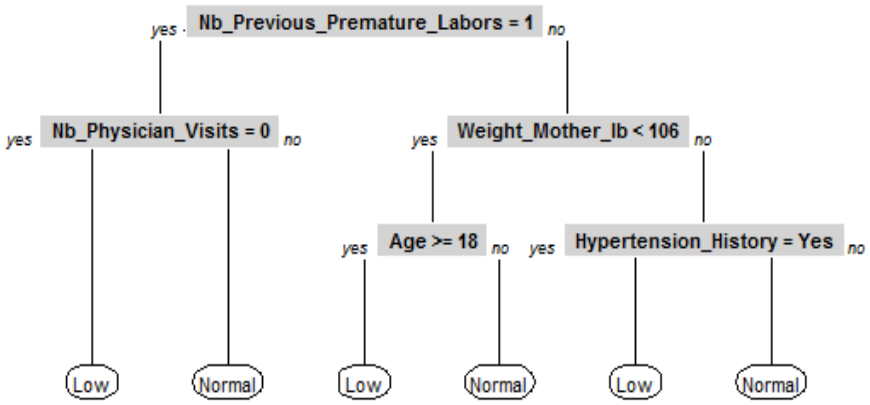


“Black Box” Machine Learning Methods

- We will illustrate the use of
  - Random forests
  - Gradient Boosting Machines

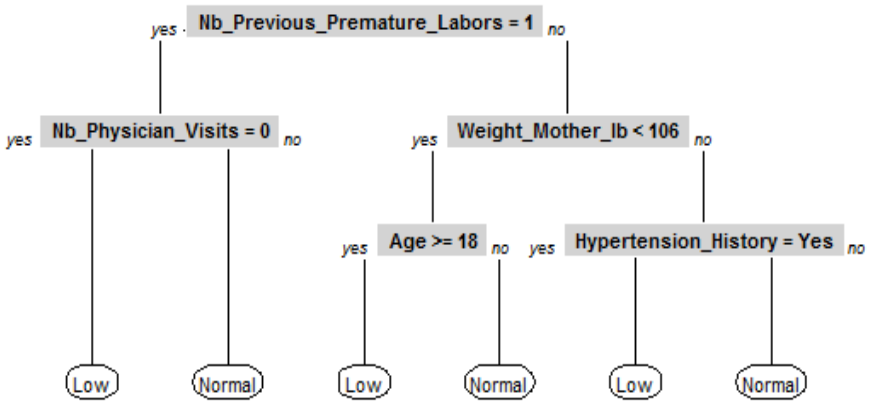
… both use decision trees as building blocks, so first some background on CART

Decision Trees, CART



- CART – Classification and Regression Trees
- Visualized as decision graphs with nodes, splits, branches, leaves.
- Each branch culminates at a leaf node and defines a region in the p-dimensional input space.
- A leaf node (region) is assigned a predicted outcome – a numeric constant for regression or a class label for classification.
- Interactions between variables are naturally represented along each branch and learnt from data.

Decision Trees, CART



- CART – Classification and Regression Trees
- Visualized as decision graphs with nodes, splits, branches, leaves.
- Each branch culminates at a leaf node and defines a region in the p-dimensional input space.
- A leaf node (region) is assigned a predicted outcome – a numeric constant for regression or a class label for classification.
- Interactions between variables are naturally represented along each branch and learnt from data.

- Learning a tree-based model involves several steps and various methods are available for each:
  - How to choose splits at each node
  - How to decide when splitting should stop
  - How to choose the optimal size of the tree (model complexity)
  - How to assign a prediction value at each leaf
- Individual trees can be very unstable: small changes in training data may lead to trees with very different splits.
- One way of dealing with this problem is to use ensemble learning techniques

Ensemble Learning – General Idea

- Hansen and Salamon (1990) showed that predictions made by a combination of classifiers can be more accurate than predictions from a single classifier as long as each base learner is accurate and the classifiers are diverse.
- Build multiple, relatively simple prediction models - base models or learners, weak learners
- Weak learners: capable of prediction accuracy at least slightly above random guessing
- Combine weak learners into one overall model (combining their strengths), e.g., by (weighted) voting or averaging.
- Diversity can be achieved in different ways:
  - different classifiers make different errors on new data, so that if their errors are uncorrelated, the majority vote or averaging will likely lead to a correct overall classification – idea behind random forest
  - adaptively focus each new weak learner on the observations that were poorly predicted by the previous learners in the ensemble – idea behind boosting methods

Random Forests

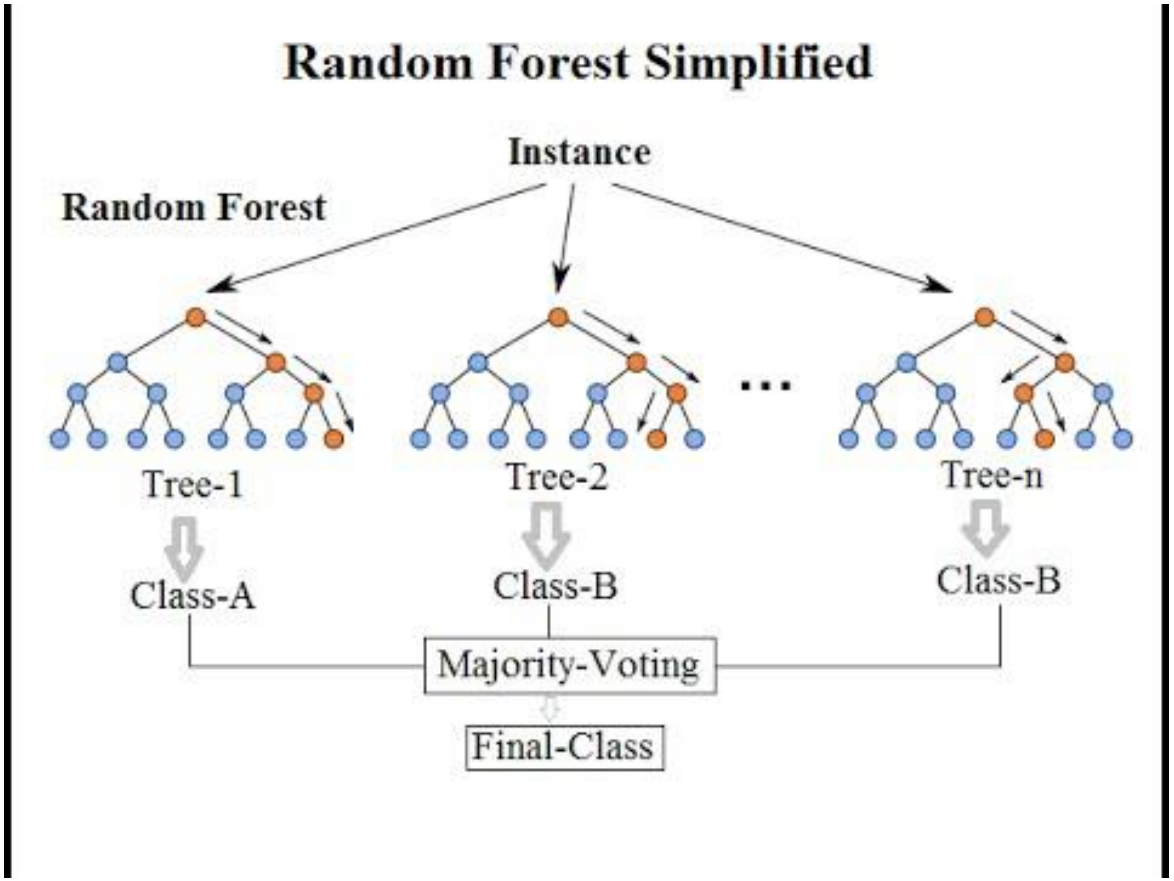


Image from: https://www.youtube.com/channel/UC95bRAbJdLyXGbMfNemr9lQ/

Random Forest

- Random forest is an ensemble learning method which builds on two ideas:
  - *bagging* (Breiman, 1996) – ensemble of unpruned trees (no restriction on the size of the trees) learned from bootstrap samples of the training data, and
  - a random feature selection during tree construction
- Bagging – a contraction of ***b***ootstrap ***agg***regation.
  - Form B bootstrap datasets from the original data and fit a model $\hat{f}^b(\boldsymbol{x}), b = 1, \ldots, B$ to each.
  - Aggregate all predictions into a single bagged prediction, e.g., $\hat{f}_{bag}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(\boldsymbol{x})$
- Why bagging improves performance of base learners, e.g., trees (low bias, high variance)?
  - Reduces the variance component of the generalization error.
  - Leaves the bias component unchanged, thus improves the predictive accuracy in general. Typically, to ensure the low bias, averaging is applied to full-sized (unpruned) trees.

Random Forest – Key Elements of Learning Algorithm

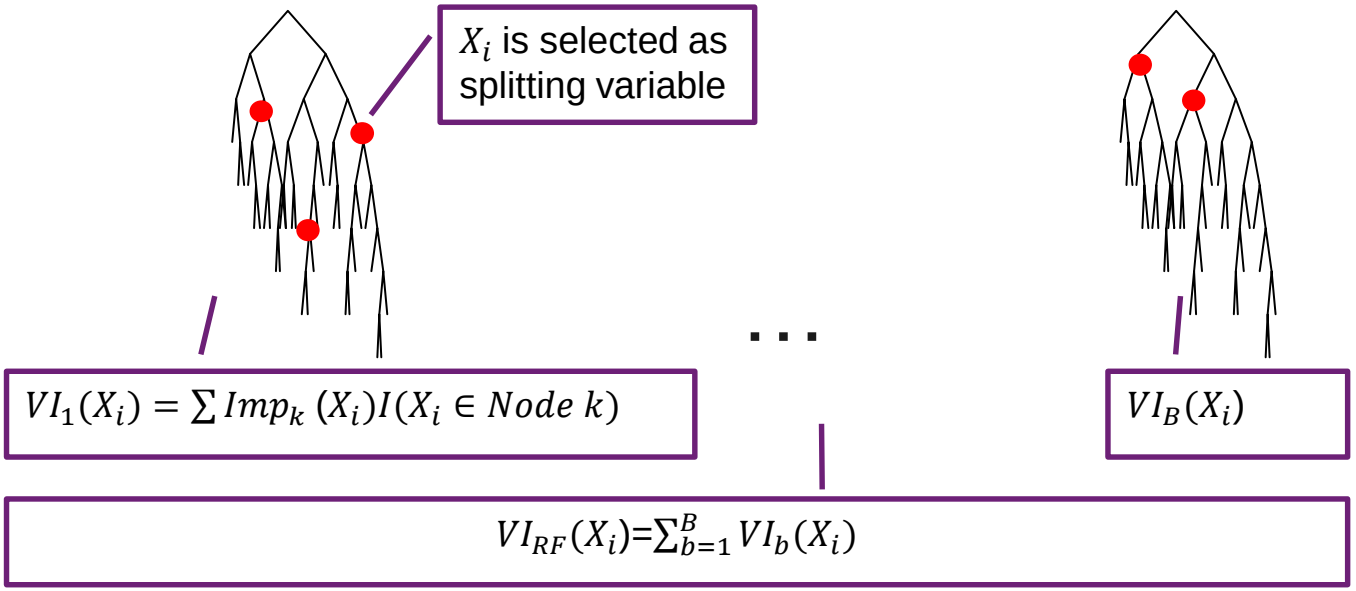
- For b = 1 to B:
  - **Draw a bootstrap sample** of the same size as the training data with replacement
  - Learn an unpruned tree on the bootstrap sample by recursively repeating the following steps until stopping criteria are met
    - **Randomly select m variables from all p predictors (m < p)**
    - Find the best split candidate from those m variables
    - Split the node into two children nodes with the best candidate
  - Combine the predictions of B trees
    - Majority vote or average probability for classification.
    - Average for regression.

Random Forests - Advantages

- More accurate than an individual tree model (tree assumes a piecewise constant true model with non-smooth boundaries which is rarely the case)
- Runs efficiently on large data sets, including many predictors and interactions
- Unbiased estimate of the generalization (prediction) error based on out-of-bag error evaluation
- Provides variables importance scores
- Some interpretability can be achieved by estimating the marginal effect of a variable (or interaction) on outcome using low-dimensional visualizations

Random Forests – Variable Importance – Reduction in Impurity

- A variable X_i is selected for a split in a tree node if it leads to a reduction in an "impurity" criterion (e.g., Gini index for classification, residual sum of squares in regression).
- Variable importance of X_i in an individual tree: sum of impurity reductions over all nodes where it's used for a split
- Variable importance of X_i in a random forest: sum of variable importance values over all trees



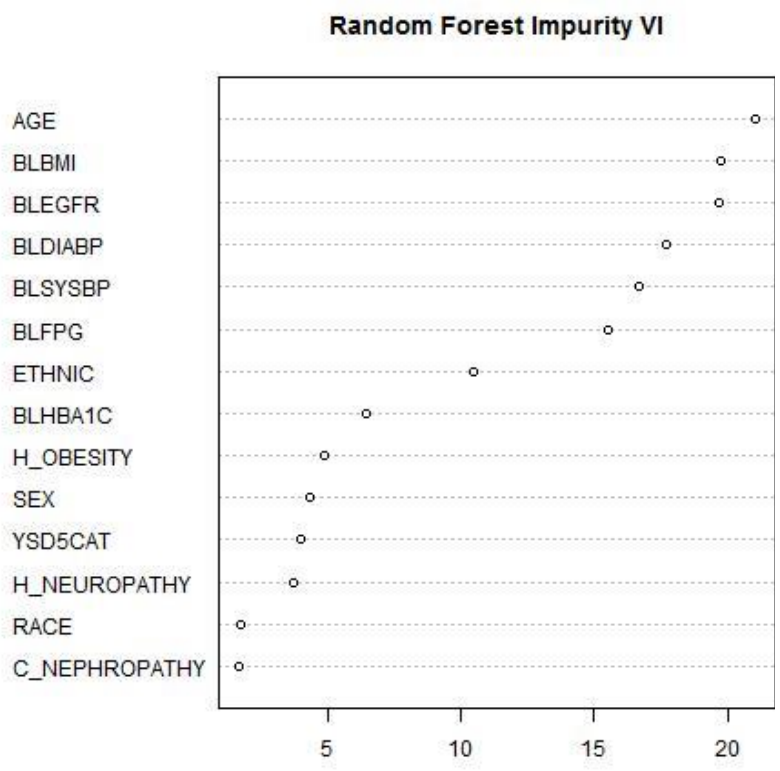
Random Forests – Variable Importance – Permutation Importance

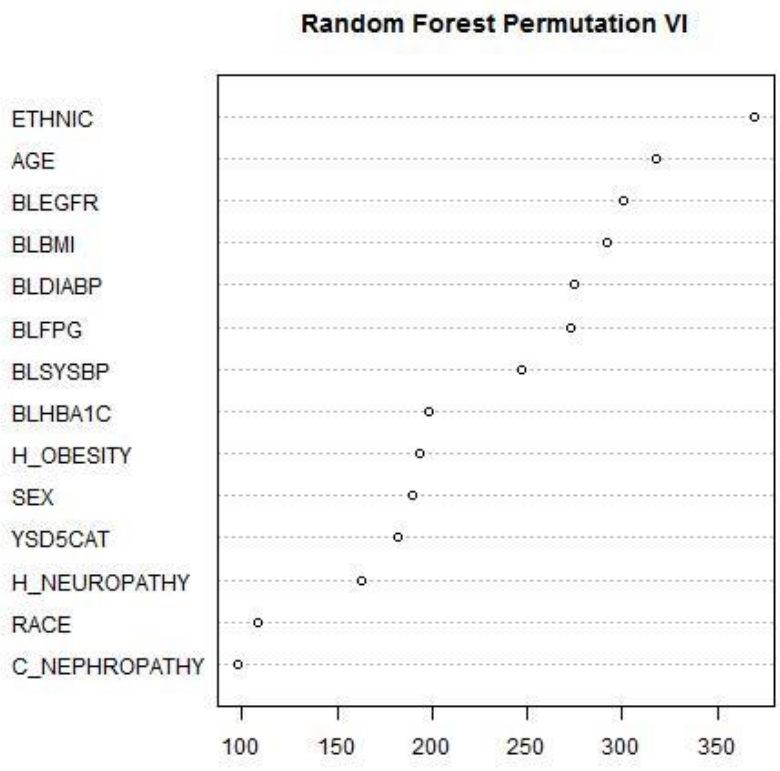
- "Permutation importance" evaluates the reduction in predictive accuracy (based on the same fitted forest) after a random permutation of the values of the variable across all training samples:
  - randomly permuting values of a variable strongly associated with response is expected to lead to substantial decreases in prediction accuracy
- Predictive accuracy, before and after permutation, is evaluated using the out-of-bag data:
  - When drawing bootstrap samples with replacement from the original data, on average, 1/3 of records are not included in a given bootstrap sample (excluded records = out-of-bag sample)
  - Each tree in a random forest is estimated from a bootstrap sample (in-bag) and evaluated for predictive accuracy on the corresponding out-of-bag sample – ensures an unbiased estimate of prediction error.
  - Note that the RF is not refit after permuting variable values - the same RF model (i.e., without refitting) is used to predict values after permutation

Random Forests – Impurity and Permutation VI for Example Dataset

R package "randomForest"

```
rf<-randomForest(pred.matrix.train, as.factor(y.train), ntree=50000, importance=TRUE)
VI.Perm=importance(rf, type=1) ### type=1: Permutation VI
VI.Impur=importance(rf, type=2) ### type=2: Impurity VI
```





Random Forest – Dealing with Variable Selection Bias

- Selecting a variable and a value for a split: CART performs an exhaustive search over all possible variables and their values for splits optimizing a measure of node impurity and selects the best variable/split value.

- Variable selection bias: favoring covariates with many possible splits – continuous or categorical with many categories.

- **Variable importance based on impurity reduction in CART inherits variable selections bias**

- Conditional trees select variables in an unbiased manner by estimating a regression relationship using binary recursive partitioning in a conditional inference framework. A splitting variable is selected first, independently of the splitting value (without an exhaustive search over all splits for all variables).

- Random forest with conditional trees is available in the R package “party”.

Random Forest – Dealing with Correlated Predictors – Conditional Variable Importance

- Permutation VI is not robust when predictors are correlated.
  - Permutation for X_i is done permuting values of X_i against outcome Y and remaining predictors $Z = X_1, \ldots, X_{i-1}, X_{i+1} \ldots . X_p$
  - High variable importance may be caused either by breaking the true correlation between X_i and Y or by breaking a spurious correlation between X_i and Y induced through correlation between X_i and some of the true predictors of Y in Z.
    - Correlated predictors “artificially” appear more important than uncorrelated ones.

Random Forest – Dealing with Correlated Predictors – Conditional Variable Importance

- Conditional VI (Strobl et al., 2008): the goal is to measure importance in the spirit of conditional correlation, i.e., measure association between X_i and Y given a correlation structure between X_i and Z:
  - X_i is permuted only within groups of records with $Z = z$ in order to preserve the correlation structure. I.e., if X_i is related to Y through Z, this association is preserved in the permuted data
  - Permutation is done within a grid defined by (a subset) of variables and is tree-specific
  - Define grids based on the partition of the predictor space induced by a tree – it has already been identified as part of tree learning
  - A subset of variables to be conditioned on (to define the grid) can be chosen as those that have a correlation with X_i that is larger than a threshold.

Y	X_j	Z
y_1	$x_{\pi_j(1),j}$	z_1
$\vdots$	$\vdots$	$\vdots$
y_i	$x_{\pi_j(i),j}$	z_i
$\vdots$	$\vdots$	$\vdots$
y_n	$x_{\pi_j(n),j}$	z_n

Y	X_j	Z
y_1	$x_{\pi_{j\|Z=a}(1),j}$	$z_1 = a$
y_3	$x_{\pi_{j\|Z=a}(3),j}$	$z_3 = a$
y_{27}	$x_{\pi_{j\|Z=a}(27),j}$	$z_{27} = a$
y_6	$x_{\pi_{j\|Z=b}(6),j}$	$z_6 = b$
y_{14}	$x_{\pi_{j\|Z=b}(14),j}$	$z_{14} = b$
y_{21}	$x_{\pi_{j\|Z=b}(21),j}$	$z_{21} = b$
$\vdots$	$\vdots$	$\vdots$

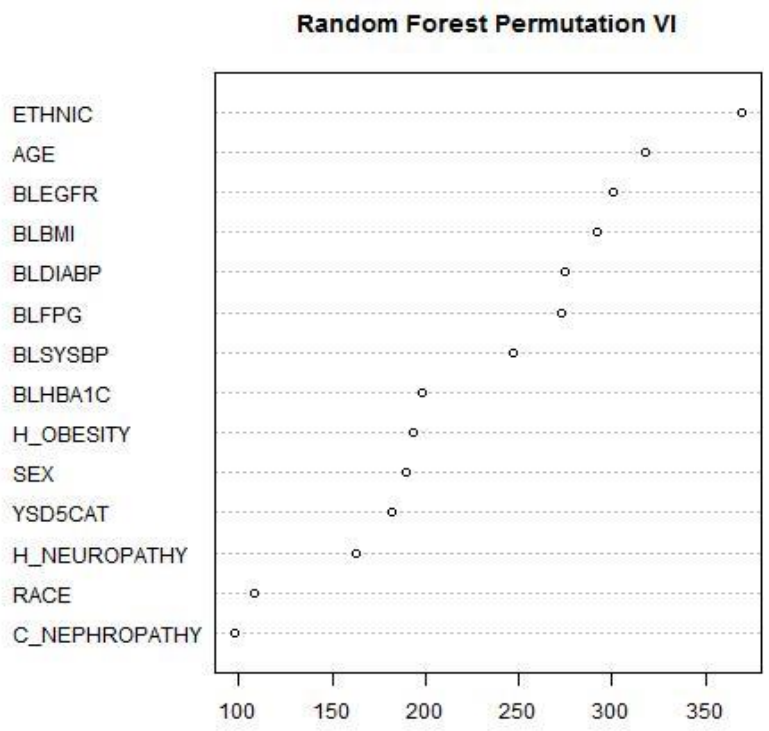
Figure 2
Permutation scheme for the original marginal (left) and for the newly suggested conditional (right) permutation importance.

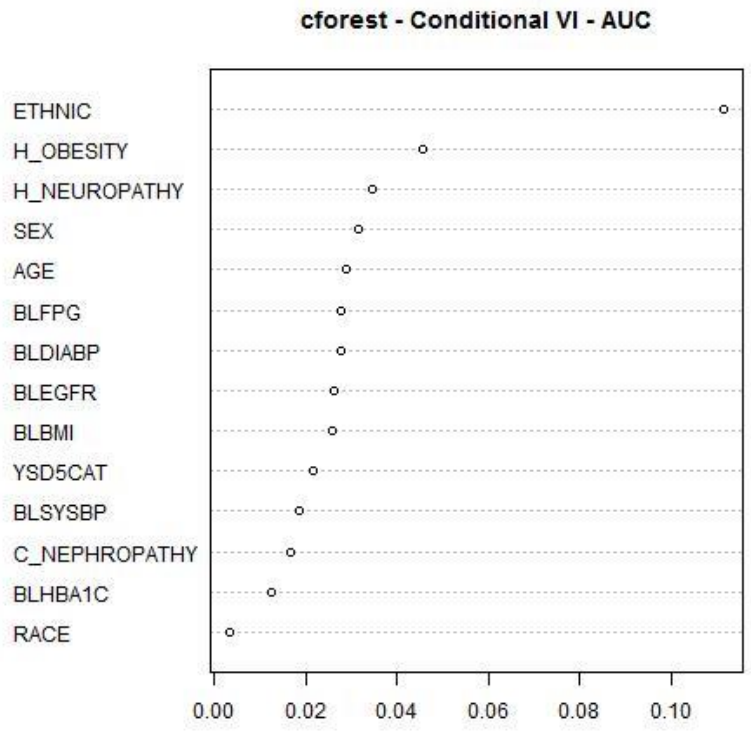
From: Strobl et al., 2008 *BMC Bioinformatics* 2008, **9**:307

Random Forests – Conditional VI for Example Dataset

R package "party"

```
my.mtry=floor(sqrt(ncol(data.train)-1))
crf<-cforest(as.formula(form.train.factor), data=data.train,
      controls=cforest_control(teststat="quad", testtype="Univariate", mincriterion=0,
      maxdepth=0, mtry=my.mtry, ntree=250, replace=FALSE, fraction=0.632,trace=FALSE))
crf.vi<-varimpAUC(crf, conditional=TRUE, OOB=TRUE, mincriterion=0.5, threshold=0.7)
```





Summary of VI Measures in Random Forest

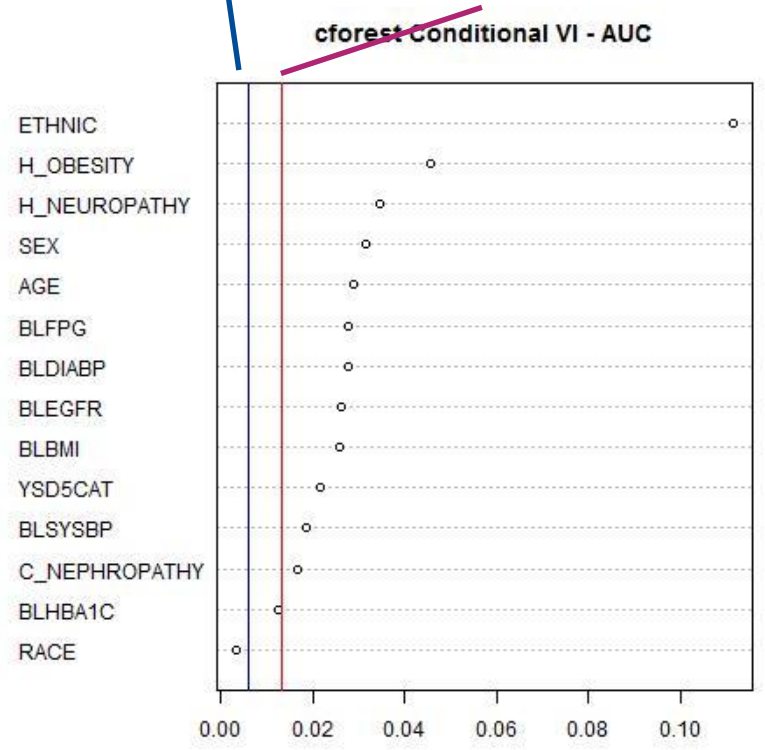
- If predictors are all of the same type and uncorrelated – use either impurity or permutation VI
- If predictors are not all of the same type but uncorrelated – use permutation VI
- If predictors are correlated – use conditional VI
- Examine ranking of variables when running with different random seeds. If it changes – increase the maximum number of trees in the random forest.

Random Forests – Global Null Test for Conditional VI

- VI scores are relative measures, but how do we know any predictors are important?
- Test a global null hypothesis of no predictor effect
  - Permutation test: For each $b = 1, \ldots, B$
    - permute outcomes, $y_{p(n)}$, against subjects' covariate values $\boldsymbol{x}_n$, where $p(n)$ is a random permutation of integers $1, \ldots, N$;
    - estimate "cforest" based on null (permuted) data;
    - compute maximum conditional VI score, $maxCVI_{0b}$, over all predictors
  - Compute p-value as the proportion of $maxCVI_{0b}$ scores which are larger than $maxCVI_{orig}$
  - Other useful summaries of the empirical null distribution:
    - Threshold= $\overline{maxCVI_0} + k\, Var(maxCVI_0)$, where $k = \Phi^{-1}(1-\alpha)$
    - 95% Upper Confidence Limit (UCL) from the empirical distribution
- For the example dataset, global test p-value< 0.001

$\overline{maxCVI_0} + k\, Var(maxCVI_0)$

95% UCL of $maxCVI_0$ (permutation-based)



Boosting

- The weak learner (e.g., a short tree – CART with small depth, often a "stump") is applied M times to the modified – re-weighted – training data sets.
- A basic idea:
  - At iteration $m = 1$, a classifier $\hat{f}_1(\boldsymbol{x})$ is estimated on data samples with equal weights $w_i = 1/N$.
  - At subsequent iterations, $m = 2, 3, \ldots, M$, weights are increased for those observations that were misclassified by the classifier from the previous iteration, $\hat{f}_{m-1}(\boldsymbol{x})$, and decreased for observations that were classified correctly.
  - As the algorithm progresses, successive classifiers focus on "difficult" cases missed by previous classifiers.
  - The overall classification is obtained as, e.g.,, $\hat{f}(\boldsymbol{x}) = sign\left[\sum_{m=1}^{M} \alpha_m \hat{f}_m(\boldsymbol{x})\right]$,

    where α_m are the weights determining the contribution of each learner based on its weighted training error.
- Boosting can dramatically increase the accuracy of very weak single classifiers (those that are just slightly better than random guessing) and outperform large single classification trees.

Gradient Boosting Machines (GBM) with Trees

- Initialize the first tree model as $f_0(\boldsymbol{x}) = \text{argmin} \sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}))$, where $L(y_i, f)$ is a loss function, e.g., exponential loss for binary classification (similar to binomial likelihood in logistic regression). Different loss functions are used depending on the type of outcome.
- For each $\text{m} = 1, \ldots, M$:
  - For each observation i, compute "pseudo residuals"

$$r_{im} = -\left[\frac{\partial L(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f=f_{m-1}}$$

  - Randomly select a subset of observations without replacement
  - Use the selected r_{im} as targets to fit a tree with terminal nodes R_{km}
  - Determine fitted values in terminal nodes to minimize the overall loss as

$$\gamma_{km} = \text{argmin}_{\gamma} \sum_{x_i \in R_{km}} L(y_i, f(\boldsymbol{x}_i) + \gamma)$$

  - Update $f_m(\boldsymbol{x}) = f_{m-1}(\boldsymbol{x}) + \nu \sum_k \gamma_{km} I(\boldsymbol{x} \in R_{km})$, where ν is a shrinkage parameter (controlling the speed of learning to reduce overfitting)
- Meta-parameters M and ν can be selected using cross-validation.

Exploring Interactions with GBM – Friedman's H Statistic

- Using a tree as a base learner and having varying tree depths allows us to assess the presence of interaction effects in the data.
- GBM lends itself to an ANOVA-type decomposition of the total variance associated with response into variance attributable to marginal effects, 2-way interaction effects, etc. For example,
  - if *depth=1*, only main effects can be captured and the fit can be decomposed into

$$f(\boldsymbol{x}) = \sum_j f_j\,(x_j)$$

  - if *depth=2*, two-way interactions can be captured and the fit can be decomposed into

$$f(\boldsymbol{x}) = \sum_j f_j\,(x_j) + \sum_{jk} f_{jk}\,(x_j, x_k)$$

- Freidman and Popescu (2005) developed the H statistic as a measure of interaction strength:
  - if X_j and X_k don't interact with each other, then function $f_{jk}(x_j, x_k) \equiv 0$
  - This can be teased out from respective *partial dependence (PD)* functions reflecting main effects of x_j and x_k or joint effect of x_j, x_k, *after averaging out the rest of predictors*
  - PD's will be explained further, they are denoted with capital F's, e.g. $F_j(x_j), F_k(x_k), F_{jk}(x_j, x_k)$.

Exploring Interactions with GBM – Friedman's H Statistic (Cont.)

- Under no interaction between x_j, x_k, joint PD can decomposed as a sum of PD's for main effects

$$F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k)$$

  - the H statistic is related to the fraction of variance of $F_{jk}(x_j, x_k)$ that is not captured by $F_j(x_j) + F_k(x_k)$ and ranges between 0 and 1
  - H statistics are constructed by using estimated PD (see next slides)

$$H_{jk}^2 = \frac{\sum_{i=1}^{N}\left(\hat{F}_{jk}(x_{ij}, x_{ik}) - \hat{F}_j(x_{ij}) - \hat{F}_k(x_{ik})\right)^2}{\sum_{i=1}^{N}\hat{F}_{jk}^2(x_{ij}, x_{ik})}$$

  - it can be used with interaction effects of any order
  - it is a relative measure, not comparable to testing significance of regression coefficients

- How do we know if it's significant?

Friedman's H Statistic – Significance

For binary response use a parametric bootstrap procedure to generate a null distribution of H – under no interaction effect (Freidman and Popescu, 2005)

- Generate artificial data containing only additive effects as

$$\tilde{y}_n \sim Be(f_A(\boldsymbol{x}_n))$$

  where $f_A(\boldsymbol{x}_n)$ is a closest fit with no interaction effects (e.g., estimate from GBM with trees of depth 1).

  Note that nonlinear effects can still be captured by a boosting machine with trees of depth 1 because of the sequential nature of the boosting algorithm.

- Fit a full model (e.g., allowing depth=2) to the artificial (no interaction) data and compute the H statistics.
- Repeat for many permutations $p(n)$ to obtain an empirical null distribution of H.
- Compute significance of the H statistic estimated from the original data as the proportion of null H values that are the same or larger than the observed from original data.
- Null distribution can be computed for the H statistic corresponding to each pair of variables as well as the maximum over all variables.
  - Maximum can be used for an overall test of presence of any interactions.

Exploring Interactions in Example Dataset

R package "gbm"

```
gbm.trees<-gbm(as.formula(form.train), data=useddata.train,
        distribution="bernoulli", interaction.depth = 2, n.trees=50000, shrinkage=0.001, n.minobsinnode=35,
        cv.folds=5, train.fraction=1, verbose=FALSE)
    interact.gbm(gbm.trees, useddata.train, i.var=int.vars)
```

- The p-value for the global test (maximum H statistic) = 0.004
- The most significant interactions for example dataset
  (p-values are adjusted using the Benjamini-Hochberg method for controlling a false discovery rate):
  - BLEGFR*BLDIABP: p-value < 0.001
  - BLDIABP*H_NEUROPATHY: p-value < 0.001
  - BLDIABP*BLFPG: p-value= 0.0364
  - BLDIABP*AGE: p-value= 0.0214
  - AGE*ETHNIC: p-value < 0.001
  - AGE*SEX: p-value < 0.001

Partial Dependence Plots

- Variable importance can help us assess which variables deserve attention, but we also need to understand how they are related to the outcome.

- Partial dependence plots (PDP) (Friedman, 2001): low dimensional graphs of the relationship between the outcome and a subset of predictors of interest while accounting for the average effect of other predictors in the model.

- PDPs aid in interpreting relationships represented by complex, “black box” models.
- Let $\boldsymbol{Z}_s$ be a subset of the predictors of interest (typically 1 to 3) and $\boldsymbol{Z}_c$ - its compliment.

  Partial dependence of outcome on $\boldsymbol{Z}_s$: $F_s(\boldsymbol{Z}_s) = \int f(\boldsymbol{Z}_s, \boldsymbol{Z}_c)\, p_c(\boldsymbol{Z}_c) d\boldsymbol{Z}_c$, where $p_c(\boldsymbol{Z}_c)$ is marginal density of $\boldsymbol{Z}_c$.

- Partial dependence function can be estimated from the training data by averaging out the effects of all other predictors in the model (N=number of records in training data):

$$\hat{F}_S(\boldsymbol{Z}_s) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(\boldsymbol{Z}_s, \boldsymbol{Z}_{i,c})$$

Partial Dependence Plots

- How do we get that from a "black box"?
  - We can use the black box to extract a predicted response based on any input.
  - We vary $\boldsymbol{z}_s$ on a grid of values while averaging across empirically observed values of all remaining variables $\boldsymbol{z}_c$, i.e.,

    obtain predictions for $(\boldsymbol{z}_s, \boldsymbol{z}_{n,c})$ from the estimated learner and average across all $\boldsymbol{z}_{n,c}$ in training data.

    This produces the marginal averaged response value for any desired input $\boldsymbol{z}_s$

- R package “pdp” provides an efficient implementation for up to 3-dimentional PDPs for many types of models, including random forest and boosted trees.

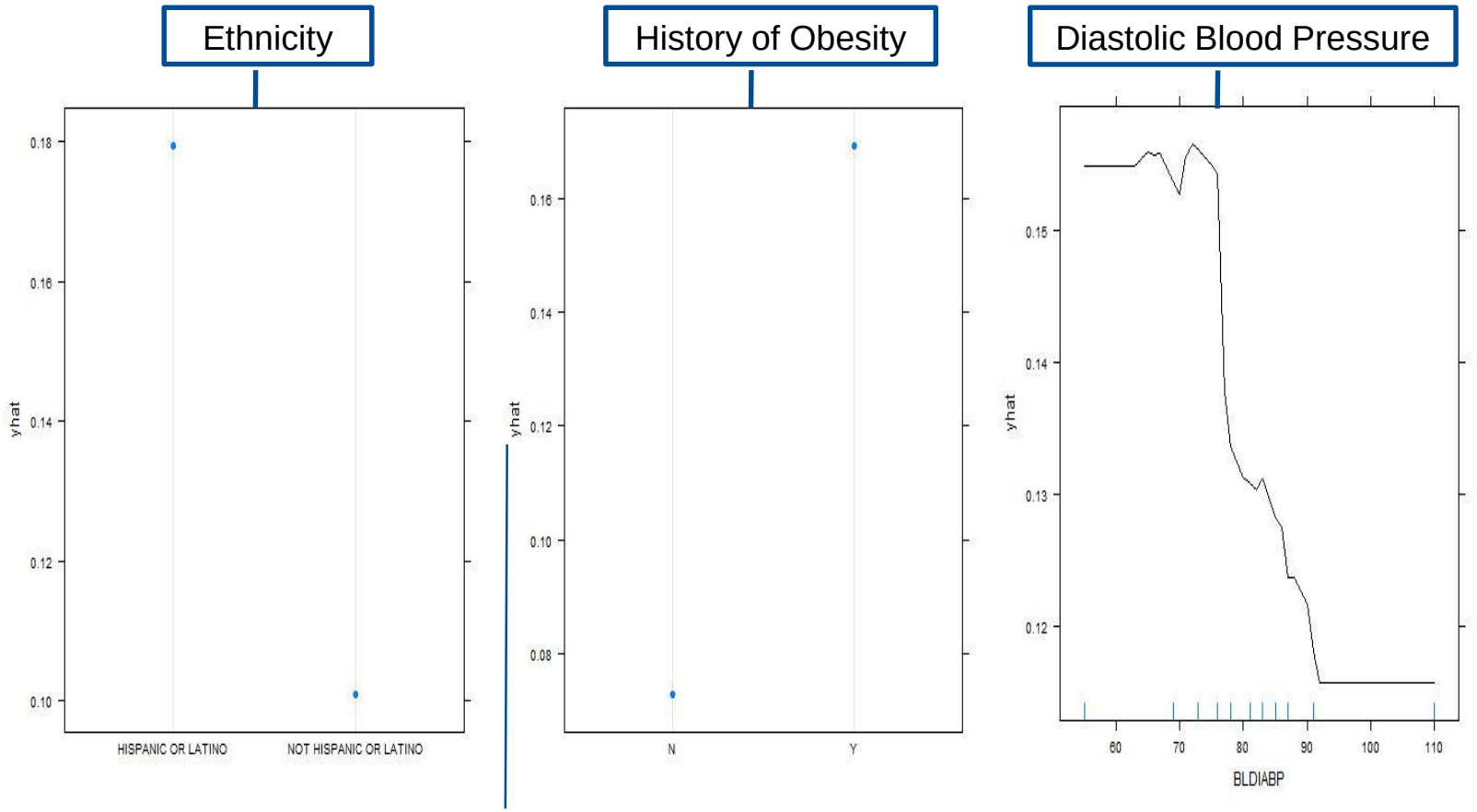
Partial Dependence Plots – “pdp” Package

R package “pdp”

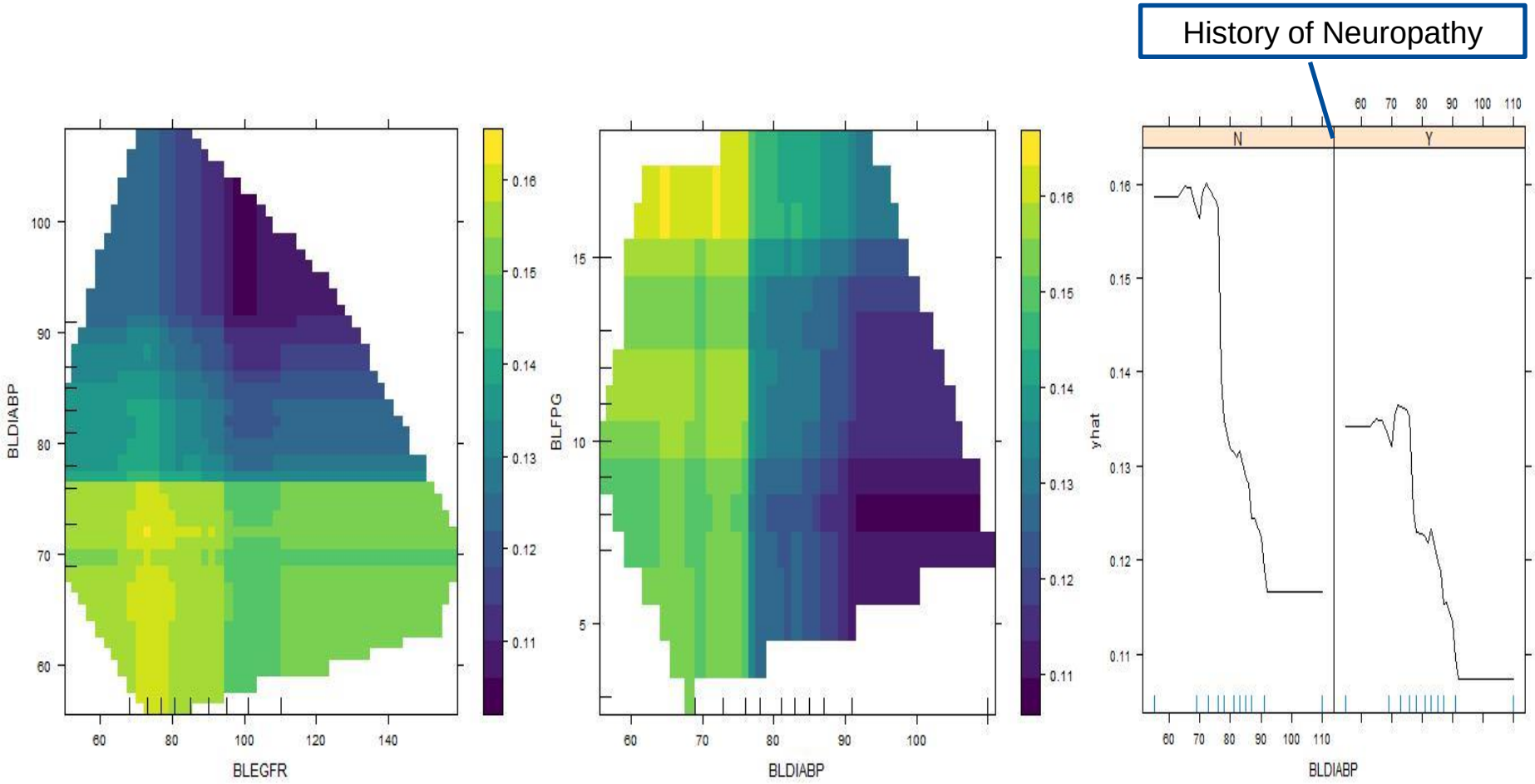
```
# Function to compute mean of predicted values in a specified dataset (to obtain plots on the probability scale)
pred.prob<-function(object, newdata){
  pred<-predict(object,newdata=newdata,type="prob")
  mean(unlist(lapply(pred, function(x) return(x[2]))))
}

partial(crf,pred.var=myvars, pred.fun=pred.prob, plot=TRUE, rug=TRUE, recursive = FALSE)
```

Specify one variable (e.g., “ETHNIC”) or a list (e.g., c(“BLDIABP”, “BLEGFR”))

Partial Dependence Plots for Example Dataset - Derived from cforest



Partial Dependence Plots for Example Dataset - Derived from cforest



Summary of What We Learned for Example Dataset

- 15% potentially avoidable dropout
- Considered 14 potential baseline predictors with 105 potential interactions
- Stepwise selection with logistic regression retained too many variables (39) to be useful, difficult to rank them in terms of importance
- Lasso also retained too many variables (64 on average), but gave some indication of the variable ordering. History of obesity, ethnicity. race, diastolic blood pressure, and history of neuropathy were at the top.
- Using ML methods, especially random forest with conditional trees, allowed us to explore a relative measure of variable importance, conditional VI being most sophisticated.
  - Global null test suggested that there were important predictors.
  - The most important variables were: ethnicity, history of obesity, and history of neuropathy.
- Gradient boosting machines allowed us to assess significant interactions
  - Global null tests suggested that there were significant interactions.
  - E.g., baseline diastolic blood pressure interacted with several other predictors.
- Partial dependence plots allowed us to visualize the associations and to assess the impact of important predictors on the probability of potentially avoidable dropout as well as approximate cut-off points for continuous predictors.

Summary of What We Learned for Example Dataset

- Based on what we learned, a targeted retention strategy may be directed towards
  - Patients of Hispanic or Latino ethnicity,
  - Patients with history of obesity,
  - Patients with normal baseline diastolic pressure, especially with
    - abnormal baseline EGFR values
    - high baseline fasting plasma glucose
    - no history of neuropathy

(Note: our example dataset is patterned after real-world studies, but not real-world data).

- Would also be useful to assess a potential impact on power using simulations and given the primary estimand and analysis method.

Proactive Retention Strategies - Caveats

- Unintentional decrease in future recruitment of subjects with a high-risk drop-out profile
- In some indications / endpoints, retention efforts directed to a subgroup of subjects may bias the endpoint outcome, e.g., in depression trials, increased subject support may influence depression rating in the supported subgroup
- Reasons for drop-out may change over time depending on external factors / current standard of care
  - Important to continuously monitor dropout rates/reasons and adjust initially designed retention strategies as needed.

Conclusions: Team Work for Better Clinical Trials

Clinical + statistical + operations + management team members should work together to:

- Learn from patterns of drop-out / missingness in previous trials
- Design trials that encourage subjects to stay all the way through without compromising wellbeing
- Statistical and machine learning methods useful in determining important predictors of potentially avoidable dropout
- Develop (targeted) retention strategies for future studies
- Promote awareness among colleagues, investigators, and site staff

References

- Breiman L (1996) Bagging predictors. Mach Learn 26:123–140
- Friedman JH (2001) Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29: 1189–1232.
- Friedman JH, Popescu BE (2008) Predictive learning via rule esembles. Ann Appl Stat. 2: 916-954.
- Hansen LK, Salamon P (1990) Neural Network Ensembles. IEEE Trans Pattern Ana. Mach Intell 12(10):993–1001
- Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Data Mining, In-ference, and Prediction, Second Edition. Springer-Verlag: New York
- Hoerl AE, Kennard R (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55–67
- Hughes S, Harris J, Flack N, Cuffe RL (2012) The statistician's role in the prevention of missing data. Pharm Stat. Sep-Oct;11(5):410-6.
- LaVange LM, Permutt T (2016) A regulatory perspective on missing data in the aftermath of the NRC report. Stat Med 35(17):2853-64
- Leon AC, Demitras H, Hedecker D (2007) Bias reduction with an adjustment for participants' intent to dropout of a randomized controlled clinical trial. Clinical Trials; 4: 540–547
- National Research Council (2010) Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The prevention and treatment of missing data in clinical trials. The National Academies Press: Washington, DC.
- Strobl C (2008) Statistical Issues in Machine Learning – Towards Reliable Split Selection and Variable Importance Measures. Dissertation, Ludwig-maximilians-universität
- Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. J R Statist Soc Series B 58(1):267-288
- Zou H (2006) The Adaptive Lasso and Its Oracle Properties. J Am Statist Assoc 101(476):1418-1429
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Stat Methodol 67(2):301-320

Questions ? Thoughts?

Bohdana.Ratitch@quintiles.com
Ilya.lipkovich@quintiles.com

Thank You!