



# Controversies in Predictive Modeling, Machine Learning, and Validation

Frank E Harrell Jr

Department of Biostatistics  
Vanderbilt University School of Medicine  
Nashville, Tennessee

Statistical Advisor, Office of Biostatistics  
US FDA Center for Drug Evaluation and Research

STRATOS 2019

Banff, Alberta

2019-06-04



# External Validation is Overrated

- Uncertainty about what is “external”
- If “external” means another time or another place, better to have a unified model with time and place
  - avoid surprises, remove temptation to label time/place differences as failure to validate
  - learn about geographical and health system differences
  - learn how to get predictions for other times and places not in dataset
- If a model is fully pre-specified, external validation validates **the** model
- Otherwise (e.g., when feature selection is used) it validates an **example** model
- Better to use resampling to validate the **process** producing the model, while being honest about instability of model selection



# Validate Researchers Instead of Models

- Many failures of research findings to replicate are predictable
- The quality of research and analysis methodology used highly influences the reliability and usefulness of the resulting research
- Validating researchers, or at least validating their analyses, is quick
- Duke Potti scandal would have been averted had Potti and Nevins shared their data and code with an independent group
  - When finally NCI obtained access, Lisa McShane obtained different results when running code twice in one day, when neither data nor code changed
- Independent research team can check reproducibility and specificity of statistical analysis plan, and can conduct their own analyses to check robustness of results



# Advantages of Bayesian Modeling

- Frequentist penalized maximum likelihood estimation works well for prediction but lack inferential methods (Hans van Houwelingen)
- Shrinkage priors with Bayes lead to plain ol' posteriors
- Sparsity priors (e.g. horseshoe) are chosen to match biological knowledge and performance goals
  - not because of availability of analytic results and fast software
- Easy to handle ordinal predictors (categorical with prior tilting towards monotonicity)
- d.f. for nonlinear effects can be data-determined and still preserve Bayesian operating characteristics



# Advantages of Bayesian Modeling, *continued*

- Instead of two-phase multiple imputation procedure, can do joint modeling of missings and outcomes
- Validation is less necessary as overfitting doesn't occur (only disagreements when the analyst used a flat prior and the reader wanted a shrinkage prior for  $\beta$ s)
- ... all the usual advantages of forward instead of backward probabilities
- E.g. compute  $P(\text{monotonicity})$ ,  $P(\text{blood pressure reduction} > 5 \text{ mmHg})$



# Advantages of Bayesian Modeling, *continued*

- Checking model assumptions distorts inference
  - PH: [Bryan Shepherd](#)
  - [On the Cost of Data Analysis](#) by Julian Faraway
- $t$ -test: prior for ratio of variances and degree of non-normality: [BEST](#)
- PH: wide priors for ordinary effects, skeptical priors for time interactions
- PO: state as polytomous logistic model with priors that shrink towards prop. odds
- Model goodness of fit **is not dichotomous**
  - acting as if it is  $\rightarrow$  false confidence if final model is interpreted as first model



# The Mirage of Variable Selection

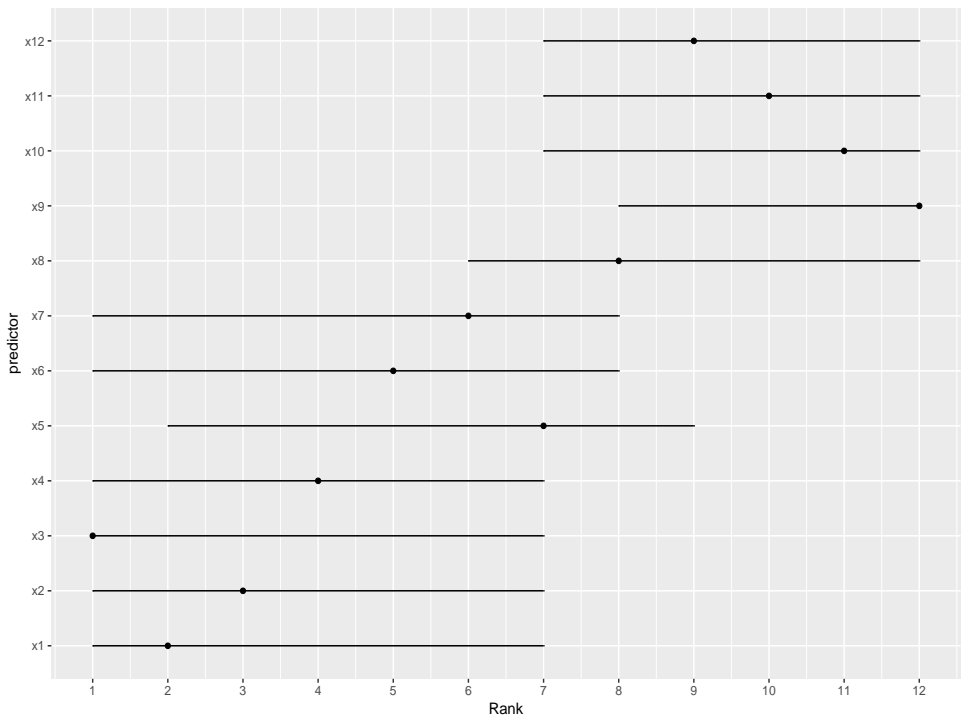
- Parsimony vs. predictive discrimination
- Feature selection requires spending information for making binary decisions that could be better used for estimation & prediction (Maxwell's demon analogy)
- $P(\text{selecting "right" variables})=0$
- Researchers worrying about FDR seldom worry about huge FNR
- Fraction of important features not selected  $\gg 0$
- Fraction of unimportant features selected  $\gg 0$



# CI for Variable Importance Quantifies Difficulty of Selection

- Bootstrap 0.95 confidence intervals for variable importance ranks
- $n = 300$ , 12 predictors,  $\beta_i = i$ ,  $\sigma = 9$ ; rank partial  $\chi^2$  (same as ranking partial  $R^2$ )



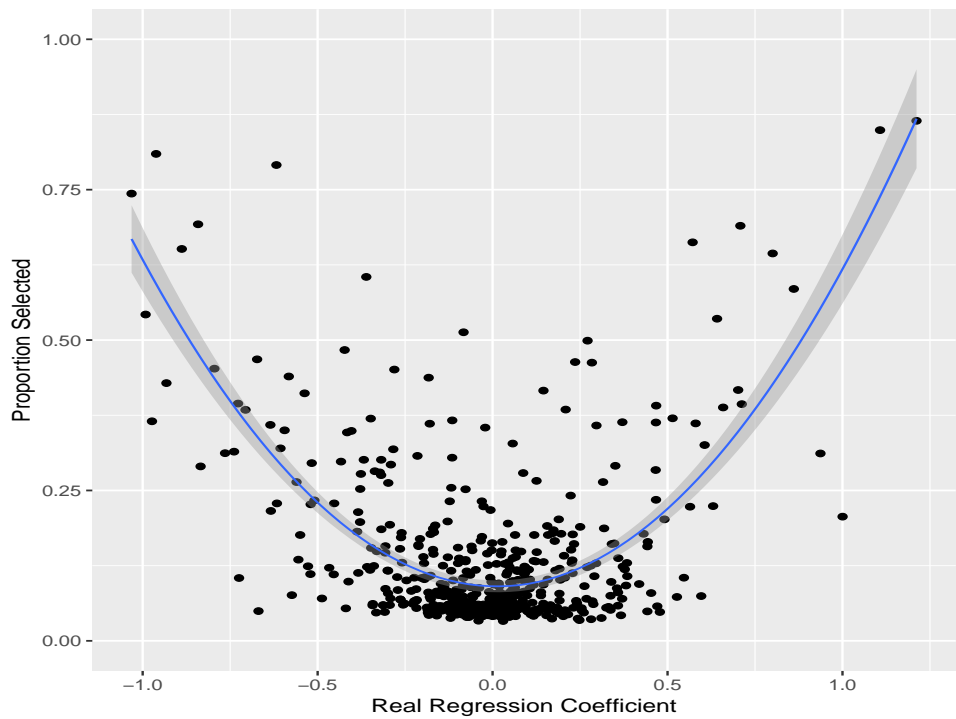




# Reliability of Feature Selection: Lasso Example

- $n = 500, p = 500, Y$  binary 0.5, all  $X$  binary 0.1, 2000 simulations
- Cross-validation on deviance used to select  $\lambda$
- $\beta$ s sampled from a Laplace distribution, giving lasso optimum performance
- $\beta$ s scaled equally to have  $c = 0.8$  for true linear predictor
- For each true  $\beta_i$  compute fraction of 2000 sims in which that variable was selected by lasso

Simulations by Shi Huang, Vanderbilt Dept. of Biostatistics





# Currently Most Stable Model Selection Method

- Assumes you actually need model selection
- Gold standard is full flexible Bayesian model with carefully chosen shrinkage (not sparsity) priors
- Project this full model onto simpler models if needed
- Piironen and Vehtari (2017): projection predictive variable selection
- [avehtari.github.io](https://avehtari.github.io) e.g. [bodyfat notebook](#)



# Machine Learning vs. Statistical Models

- Statistical models
  - Probability distribution for data
  - Favor additivity
  - identified parameters of interest
  - Inference, estimation, prediction
  - Most useful when uncertainty high
- Machine learning
  - Algorithmic
  - Equal opportunity for interactions as for main effects
  - Prediction
  - Most useful when signal:noise ratio high and sample size is huge
    - Modern modelling techniques are data hungry by van der Ploeg, Austin, Steyerberg
  - Deep learning  $\equiv$  neural network
    - neural network  $\equiv$  polynomial regression – Norm Matloff
- [fharrell.com/talk/mlhealth](http://fharrell.com/talk/mlhealth)



# The Curse of Flexibility

- Additive models:  $V(\text{effect})$  virtually constant as number of adjustment variables  $\uparrow$
- ML, recursive partitioning, etc.: synonymous with fitting high-order interactions
- Continuous  $Y$  with variance  $\sigma^2$ , sample size  $n$
- Balanced  $2^3$  design  $ab \times \alpha\beta \times 12$ 
  - 8 combinations each with sample size  $\frac{n}{8}$
- $V(\bar{Y}) = \frac{\sigma^2}{n} = v$
- $V(\bar{Y}_a) = 2v$
- $V(\bar{Y}_{a\alpha}) = 4v$
- $V(\bar{Y}_{a\alpha 1}) = 8v = V(\hat{E}(Y|a\alpha 1))$  under 3-way IA model
- $V(\hat{E}(Y|a\alpha 1)) = 4v$  under additive model



# Predictive Measures

- Gold standards
  - smooth flexible calibration curve
  - frequentist: log likelihood
  - Bayesian: log likelihood + log prior
  - explained outcome heterogeneity
  - heterogeneity of predictions (Kent & O'Quigley-type measures);  $\text{var}(\hat{Y})$
  - relative explained variation (relative  $R^2$ ): ratio of variances of  $\hat{Y}$  from a subset model to the full model
- [fharrell.com/post/addvalue](https://fharrell.com/post/addvalue)
- Majority of ML papers do not demonstrate adequate understanding of predictive accuracy
  - Recent survey of ML in medicine:  $< \frac{1}{10}$  of papers included a calibration curve



# Predictive Measures, *continued*

- Proportion “classified” “correctly”, sensitivity, specificity, precision, and recall are discontinuous improper accuracy scores
  - optimizing them will result in a bogus model
- ROC curves are highly problematic
  - coordinates: sens and 1-spec are improper scores
  - coordinates: transposed conditionals
  - invite dichotomization of predictors
  - not insightful
  - high ink:information ratio





# Predictive Measures and Decision Making

Controversies  
in Predictive  
Modeling,  
Machine  
Learning, and  
Validation

Model  
Validation

Bayesian  
Modeling

Variable  
Selection

ML and SM

Predictive  
Measures

Summary

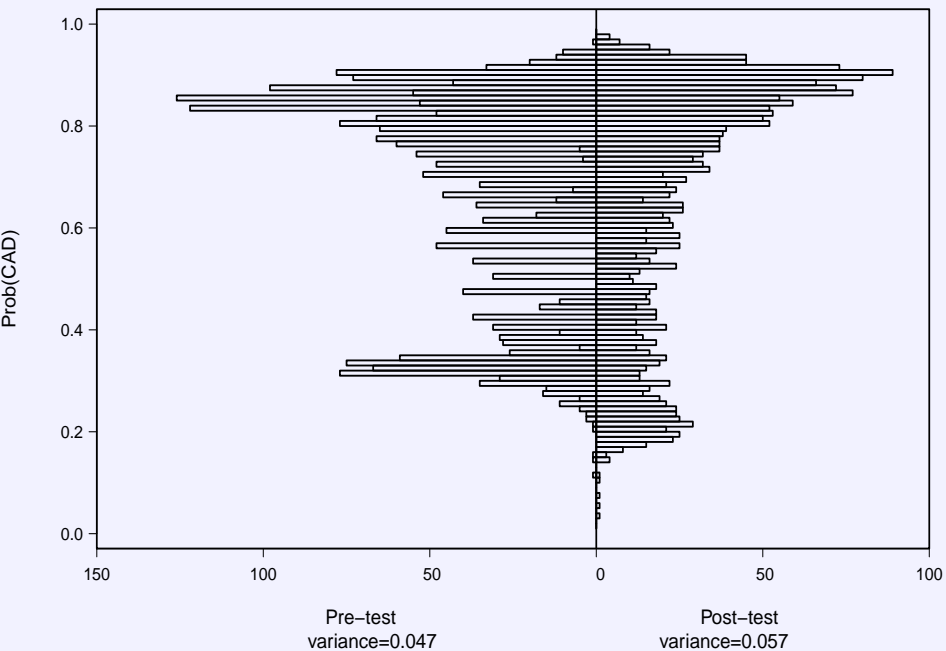
- Optimum Bayes decision that maximizes expected utility
- Expected utility uses posterior distribution of outcome probability for a patient combined with consequences of possible wrong decisions
- Measures with transposed conditionals (e.g., sensitivity) and ROC curves and AUROC (*c*-index) play no role

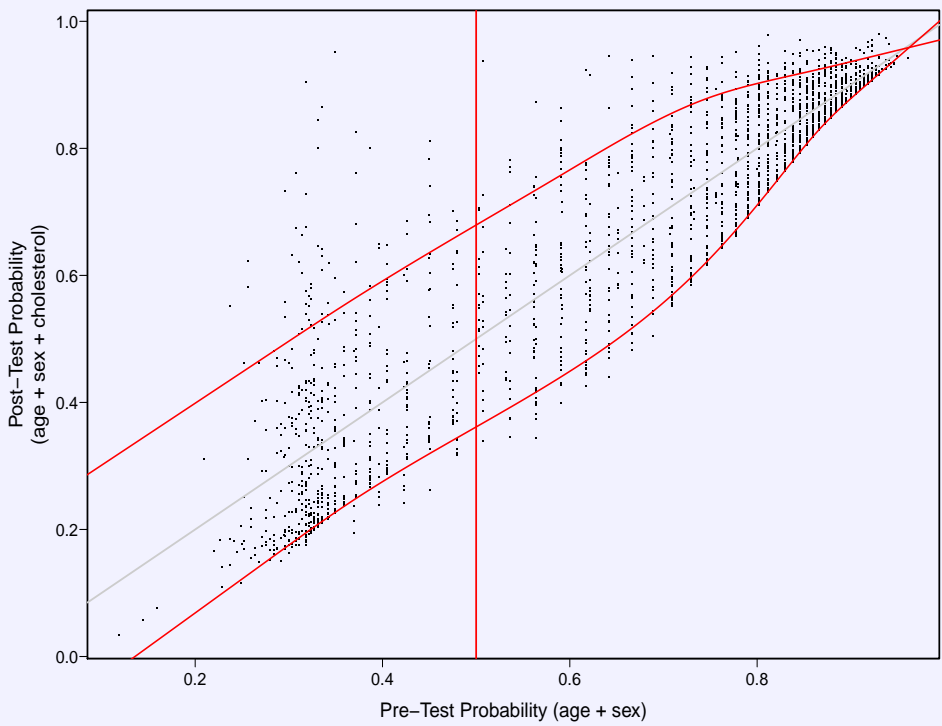


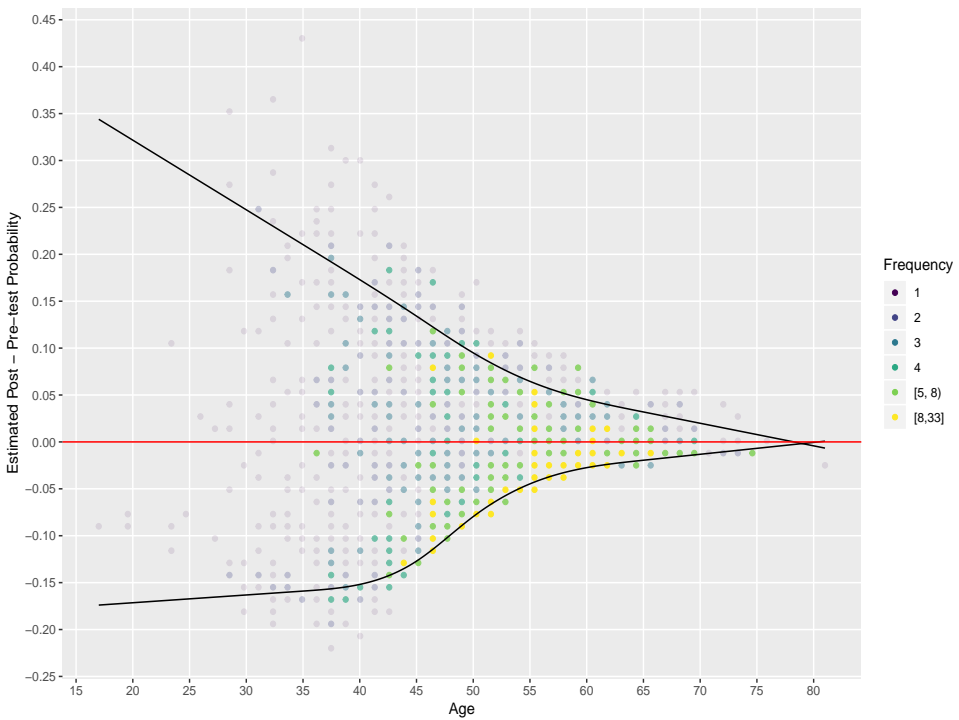
- Relative explained variation
  - ratios of  $\text{var}(\hat{Y})$
  - “Adequacy index”: ratio of model likelihood ratio  $\chi^2$
- Scatterplot of one  $\hat{Y}$  against another
- Plot differences in  $\hat{Y}$  against patient characteristics
- Example: Duke Cardiovascular Databank, patients referred for chest pain
- $Y$ : presence/absence of significant coronary disease
- Basic model:  $\text{sex} \times \text{spline}(\text{age})$
- “New” marker: total cholesterol (interacts nonlinearly with age)

Relative explained variation: 0.83

Fraction of new information: 0.17









# Summary

- Simple measures of explained variation, and gold-standard log-likelihood-based indexes have much to offer in describing model performance and comparing models
- ML and SM need to be distinguished for meaningful discussion
- Biomedical researchers are beginning to tire of ML hype and its performance in low signal:noise situations
- The information content in most datasets does not allow for reliable variable selection, and researchers are often misled
- Bayesian modeling is the way of the future
- We should spend more time validating investigators and less time doing “external” validations