VANDERBILT
School of Medicine
BIOSTATISTICS

Controversies in Predictive Modeling, Machine Learning, and Validation

Frank E Harrell Jr

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, Tennessee USA

International Conference on Recent Advances in Big Data and Precision Health
Taiwan 2022-10-03

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

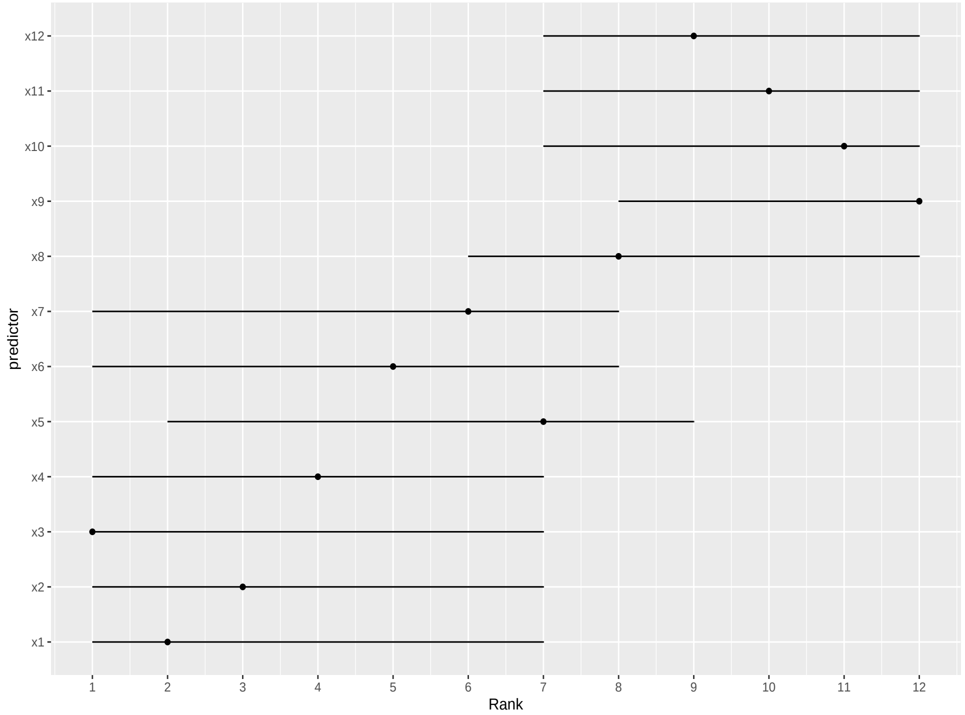
- Many failures of research findings to replicate are predictable
- The quality of research and analysis methodology used highly influences the reliability and usefulness of the resulting research
- Validating researchers, or at least validating their analyses, is quick
- Duke Potti scandal would have 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

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(selecting "right" variables)=0
- Researchers worrying about FDR seldom worry about huge FNR
- Fraction of important features not selected $>> 0$
- Fraction of unimportant features selected $>> 0$

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 χ^2 (same as ranking partial R^2)

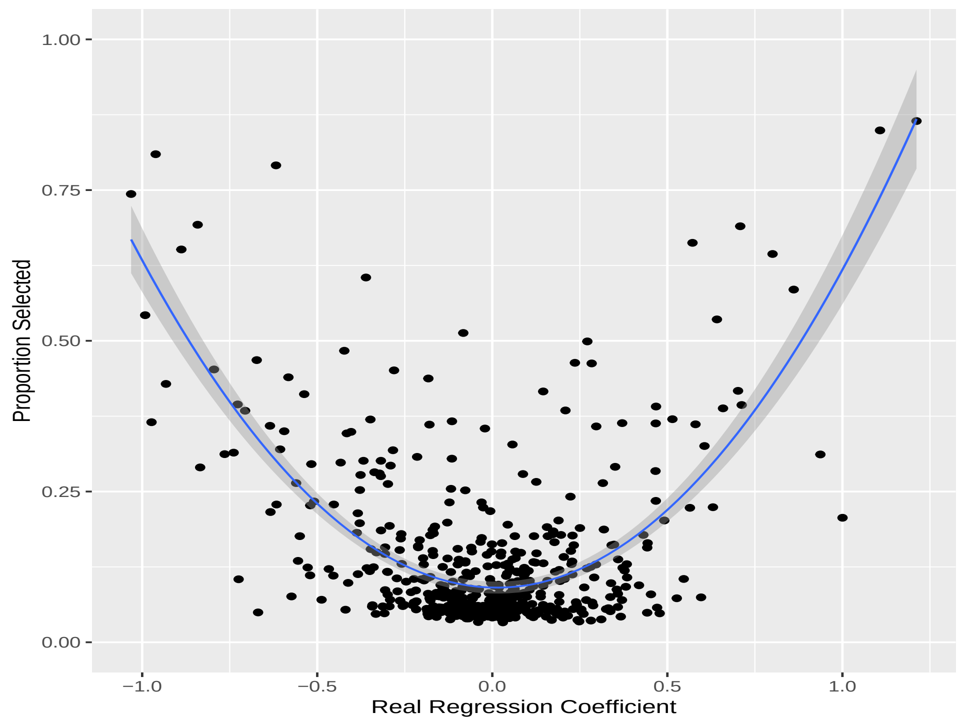

x12
x11
x10
x9
x8
x7
x6
x5
x4
x3
x2
x1
predictor
1
2
3
4
5
6
7
8
9
10
11
12
Rank

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 λ
- βs sampled from a Laplace distribution, giving lasso optimum performance
- βs scaled equally to have $c = 0.8$ for true linear predictor
- For each true β_i compute fraction of 2000 sims in which that variable was selected by lasso

Simulations by Shi Huang, Vanderbilt Dept. of Biostatistics

See also Zhao and Yu 2006 `jmlr.org/papers/volume7/zhao06a`



1.00
0.75
0.50
0.25
0.00
Proportion Selected
−1.0
−0.5
0.0
0.5
1.0
Real Regression Coefficient

Machine Learning vs. Statistical Models

- Statistical models (SM)
  - Probability distribution for data
  - Favor additivity
  - identified parameters of interest
  - Inference, estimation, prediction
  - Most useful when uncertainty high
- Machine learning (ML)
  - Algorithmic
  - Equal opportunity for interactions as for main effects
  - Prediction
  - Most useful when signal:noise ratio high
  - Deep learning $\equiv$ neural network
    - neural network $\equiv$ polynomial regression (Matloff)

Current Status: ML in Medicine

- Ultra-high dimensions (e.g., GWAS) can only be analyzed with statistical models
- Researchers usually undervalue the flexibility available with SMs
- Review articles are finding modest gains in predictive discrimination from ML when noise is high
- Majority of ML applications do not provide a calibration curve to demonstrate absolute predictive accuracy
- When they do the calibration is found to be wanting
- SMs perform quite well in most situations
- SMs are more interpretable

fharrell.com/talk/mlhealth

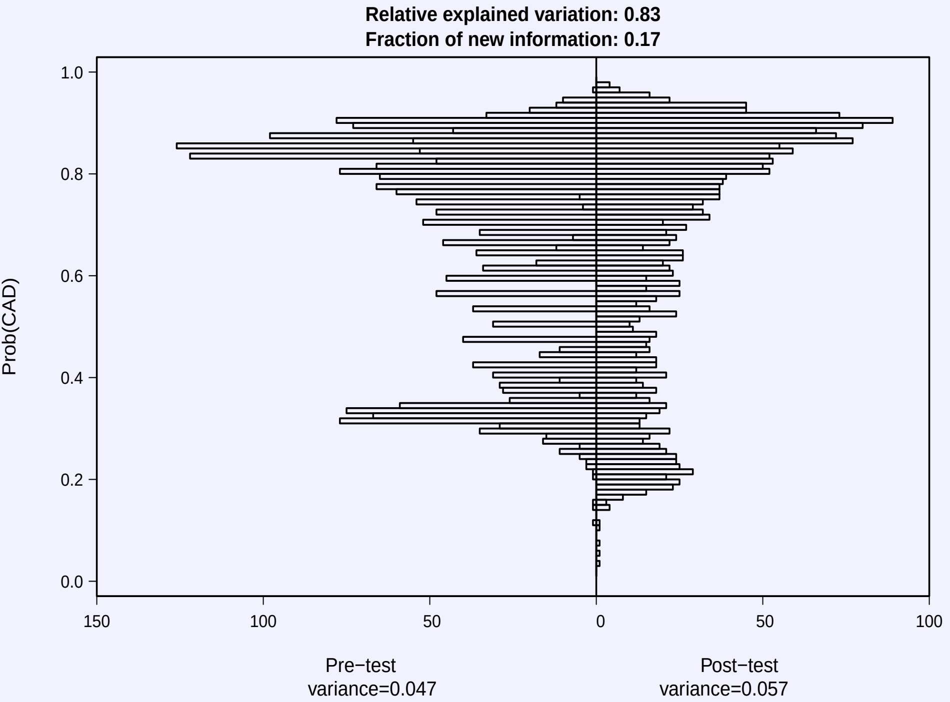
Predictive Measures and Decision Making

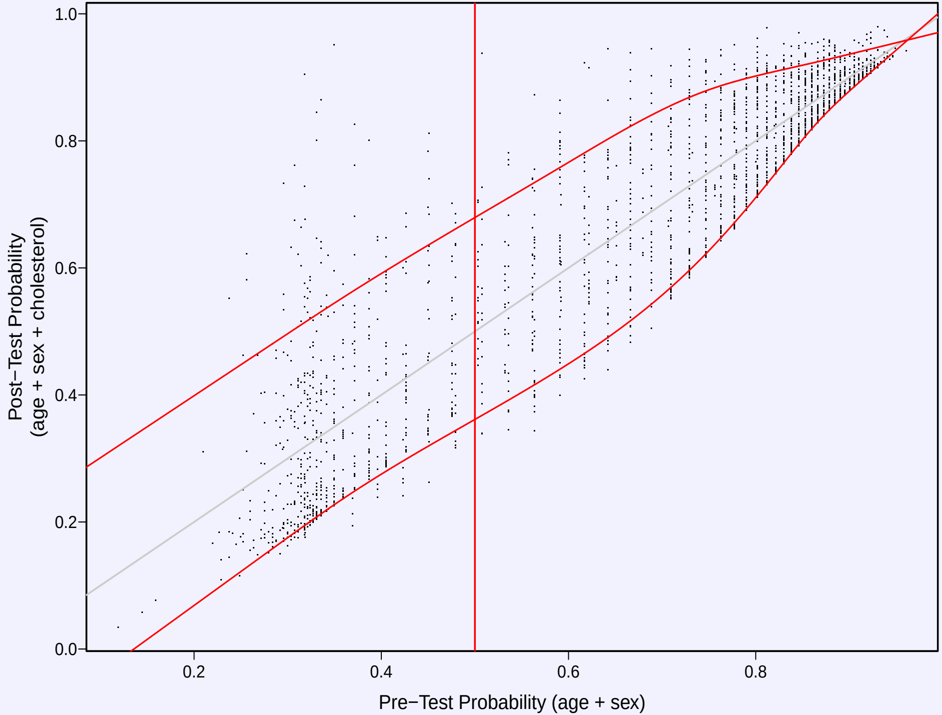
- Optimum Bayes decision that maximizes expected utility
- Expected utility uses posterior distribution of outcome probability for a patient combined with consequences of possible wrong decisions
- Measures with transposed conditionals (e.g., sensitivity) and ROC curves and AUROC (c-index) play no role

Quantifying Predictive Information

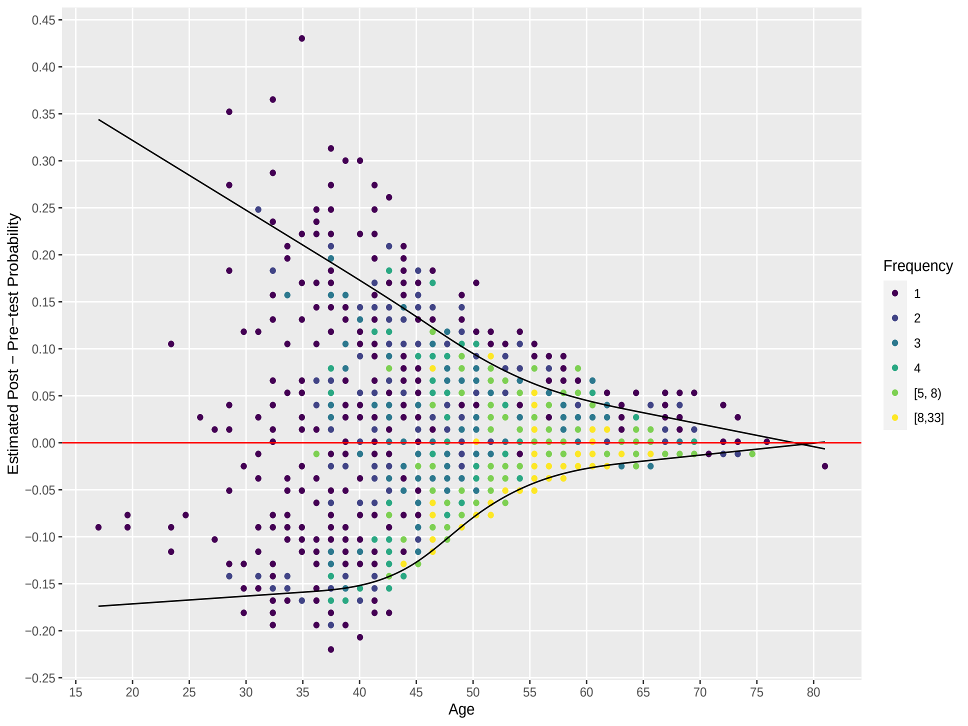
- Relative explained variation
  - ratios of var($\hat{Y}$)
  - "Adequacy index": ratio of model likelihood ratio χ^2s2
- Scatterplot of one $\hat{Y}$ against another
- Plot differences in $\hat{Y}$ against patient characteristics
- Example: Duke Cardiovascular Databank, patients referred for chest pain
- Y: presence/absence of significant coronary disease
- Basic model: sex $\times$ spline(age)
- "New" marker: total cholesterol (interacts nonlinearly with age)

fharrell.com/post/addvalue


Relative explained variation: 0.83
Fraction of new information: 0.17
Prob(CAD)
1.0
0.8
0.6
0.4
0.2
0.0
150
100
50
0
50
100
Pre−test
variance=0.047
Post−test
variance=0.057



Post−Test Probability
(age + sex + cholesterol)
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
Pre−Test Probability (age + sex)



Frequency
1
2
3
4
[5, 8)
[8,33]
Estimated Post − Pre−test Probability
Age
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
15
20
25
30
35
40
45
50
55
60
65
70
75
80