Development and Validation of a machine learning based prediction model in a clinical trial setting

Eric Polley

Biomedical Statistics and Informatics
Mayo Clinic
Polley.Eric@mayo.edu

HSR Grand Rounds
Mayo Clinic
March 2019

Clinical Prediction Models

Three types of predictors:

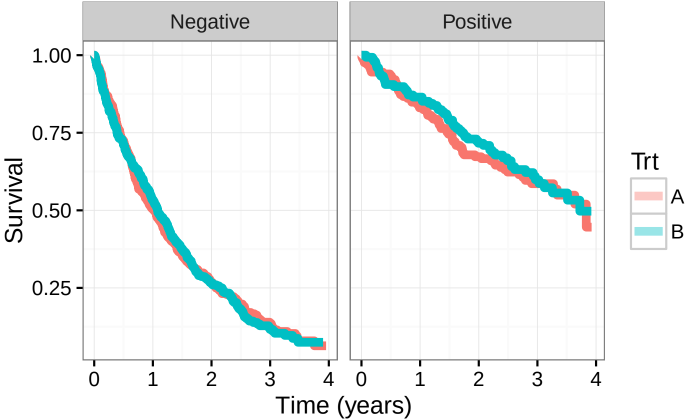
- Diagnostic
- Prognostic
- “Predictive” or treatment selection

Diagnostic

Diagnostic models often built with a binary (*e.g.* prevalent disease yes/no) but can also incorporate disease subtypes as a multiclass outcome.

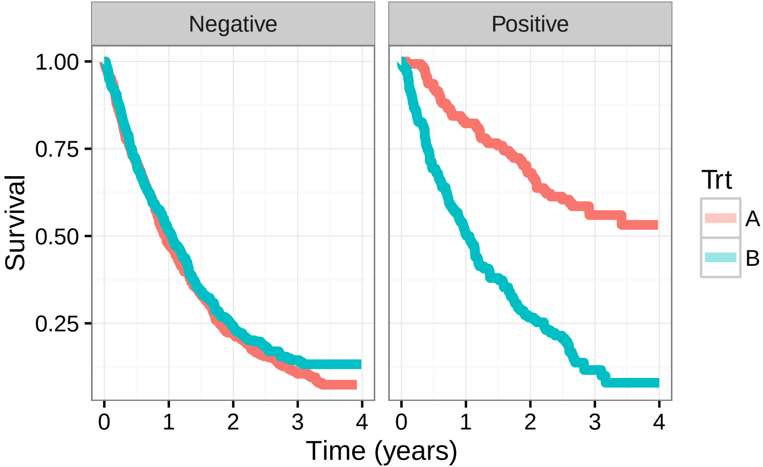
Prognostic

Example of a prognostic, but not predictive risk score:



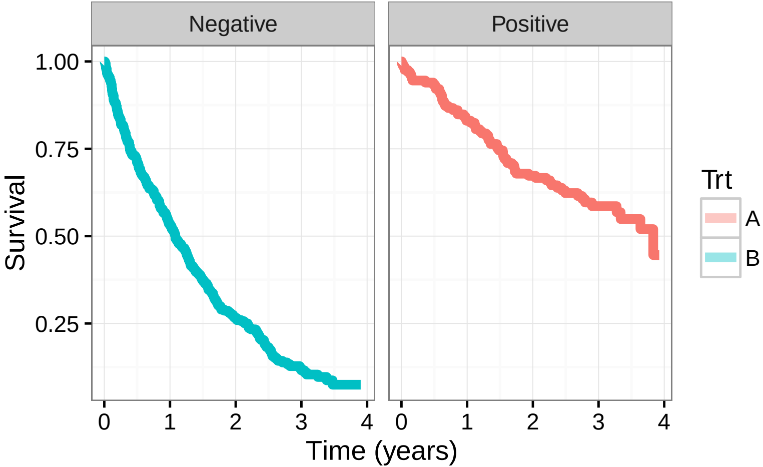
Predictive

Example of a predictive, but not prognostic risk score:



Prognostic or Predictive

If study uses the risk score to select treatment

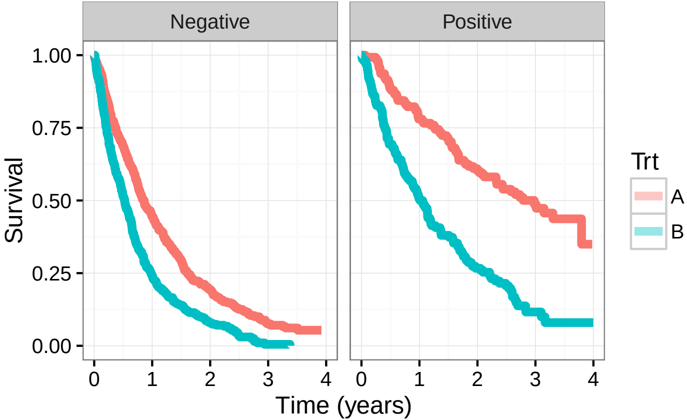


Predictive

Predictive risk scores often simplified as treatment by clinical feature interaction in a statistical model, but this isn't sufficient

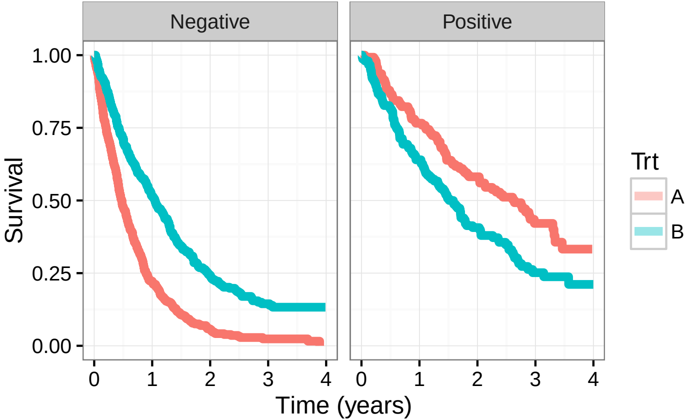
Predictive Risk Score

Treatment by risk score, same direction of effect:



Predictive Risk Score

Treatment by risk score, different direction of effect:



Estimation

A framework for the estimation and testing of a predictive model is the "Adaptive Signature Design" by Freidlin and Simon[1]

The goal of the framework is to identify a predictive model based on baseline covariates to predict patients likely to be respond better to the experimental arm (arm E) relative to the control arm (arm C)

[1] https://www.ncbi.nlm.nih.gov/pubmed/16278411

Adaptive Signature Design

Define the following for the predictive modelling:

- t_i is the randomized treatment assignment (1 for arm E, 0 for arm C)
- $x_{1i}, \ldots, x_{mi}$ are the m features
- p_i is the probability of response
- $\log \frac{p}{1-p} = f(t, x|\beta)$ is the parametric model predicting treatment response
- $f(t, x|\beta) = \mu + \lambda t + \eta_1 x_1 + \ldots + \eta_m x_m + \gamma_1 t x_1 + \ldots + \gamma_m t x_m$
- λ and η are the main effects for treatment and the features
- γ are the interaction effects
- $exp(\lambda + \gamma x_i) > R$ used to identify patients likely sensitive to E

Adaptive Signature Design

The trial is designed to enroll N patients with equal randomization to treatment arms

- The sample is split into two stages, with $N_1 + N_2 = N$
- An overall test for treatment effect (unadjusted) is made using all N patients at level α_1
- Using only N_1 patients, estimate the parameters in $f(t, x|\beta)$
- Classify the N_2 patients from stage 2 as sensitive or not, and perform a test of treatment effect only within the subset predicted to be sensitive at level α_2
- Study is statistically significant with either test rejects the null hypothesis and the overall significance level is controlled at $\alpha = \alpha_1 + \alpha_2$.

Adaptive Signature Design

In the original paper, the authors recommended $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$ on the assumption the sensitive subpopulation was likely small but had a large treatment effect.

Requires a larger sample size than an overall phase III clinical trial testing only the overall effect, but allows the identification and testing of a predictive signature in cases where the sensitive subpopulation isn't known *a priori*.

If predictive signature already available, enrichment design is recommended instead.

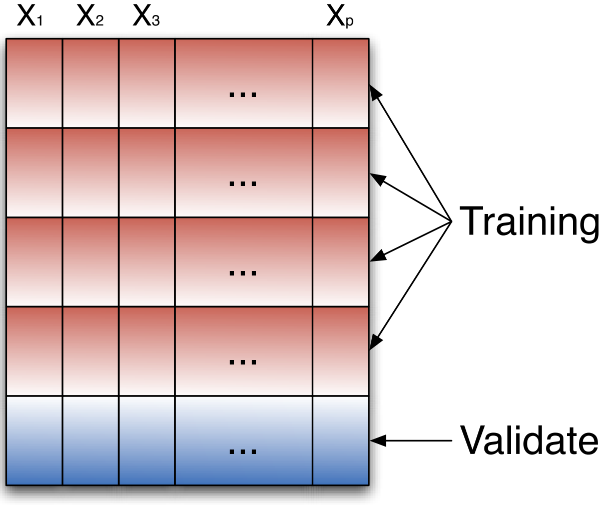
Cross-Validation

The framework was extended to incorporate V-fold cross-validation in 2010[2]

While statistically valid, the previous framework was inefficient because of the sample split process. Updated proposal demonstrates how K-fold cross-validation can be utilized in a trial setting to improve efficiency.

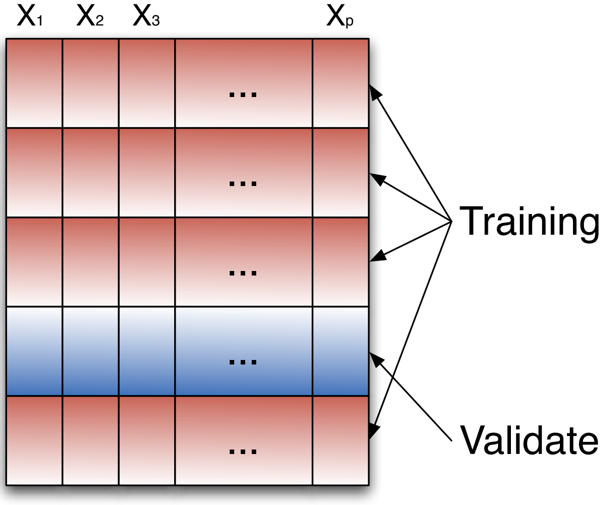
[2] https://www.ncbi.nlm.nih.gov/pubmed/20068112

K-Fold Cross Validation

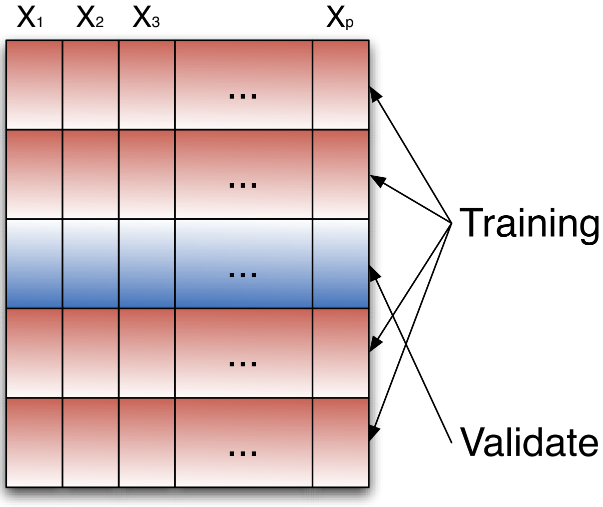


MAYO CLINIC

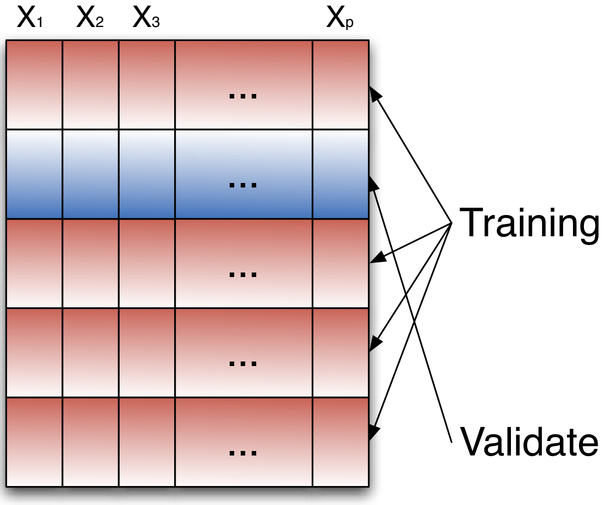
K-Fold Cross Validation



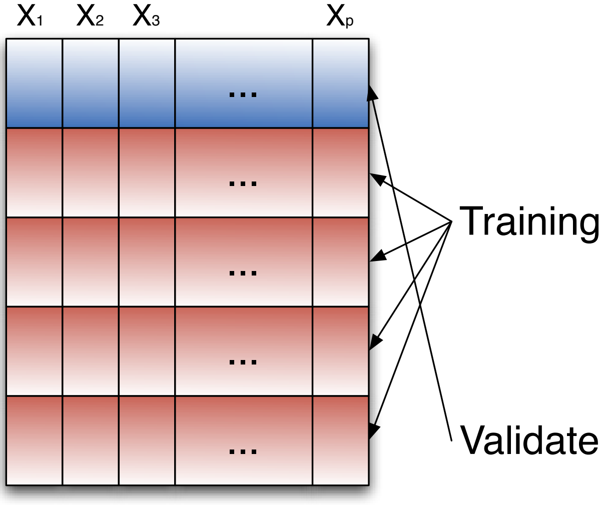
K-Fold Cross Validation



K-Fold Cross Validation



K-Fold Cross Validation



MAYO CLINIC

CV Adaptive Signature Design

Utilizing similar notation from above, but add

- Randomly partition N patients in K mutually exclusive blocks
- Define V_k to be the set of patients in the k^{th} validation block, and the remaining patients in D_k, the development set
- For each D_k, estimate the parameters in the predictor and apply in the corresponding V_k set.
- Stack the predicted sensitivity classifications across all K folds and estimate the test statistic, T, for the subgroup treatment effect.
- Since CV creates a correlation structure, the authors recommend the permutation test to obtain a p-value

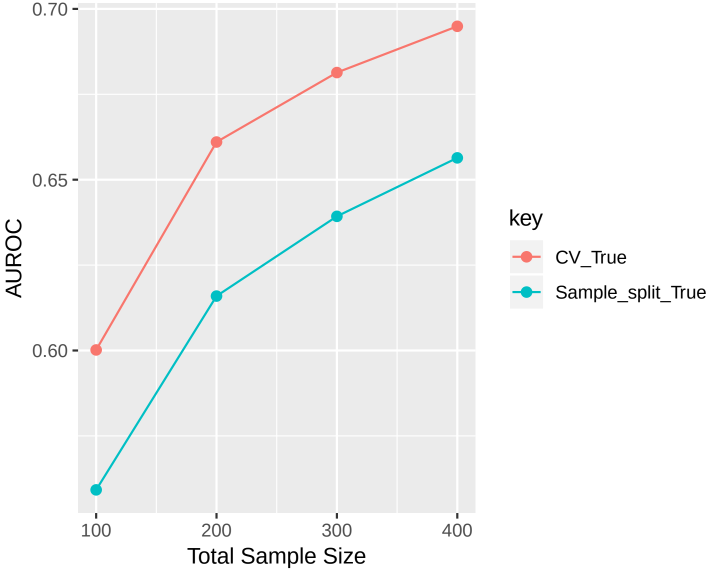
Permutation Test

- For the permutation test, shuffle the treatment values, t in the full dataset
- Repeat the entire cross-validation process as above on the observed data
- Estimate the test statistic with the permuted data, T^*
- Repeat the permutation procedure B times
- The p-value is $\frac{1+\sum \mathrm{I}(T^* \geq T)}{1+B}$
- Requires all steps for model development to be repeated within the permutations and cross-validation steps

CV Adaptive Signature Design

- If significant, the final predictor is trained on the entire N samples
- Most algorithms include a cross-validation step to select tuning parameters, this must be done nested within an additional inner cross-validation procedure
- In a simulation with 30% sensitive, show an increase in power from 0.589 to 0.641 by adding CV

Cross-Validation



Cross-Validation

Cross-Validation has a few assumptions to be a valid estimate of predictive performance

Cross-Validation

- With $N = 100$, and $P = 6000$, and binary outcome ($M = 2$)
- Simulated dataset assuming independent multivariate normal for the data generating distribution.
- Built a classifier with Linear Discriminant Analysis on the subset of genes differentially expressed at $\alpha = 0.001$ level.

Cross-Validation

Resubstitution method

- Use all 100 samples to build LDA classifier
- Evaluate misclassification error rate on the same samples

Leave one out cross validation (LOOCV) without gene selection

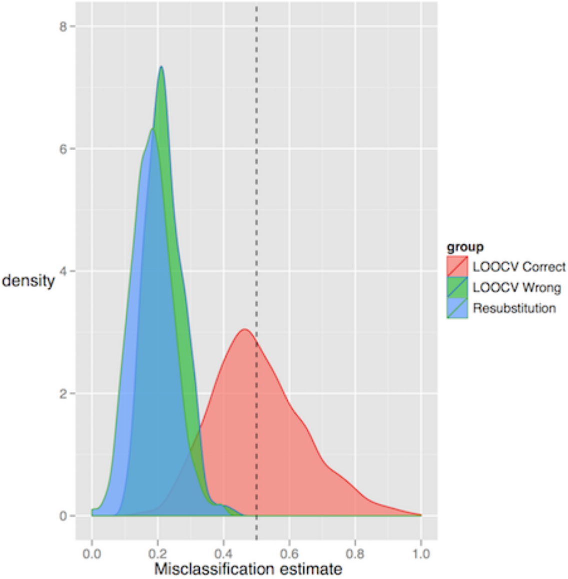
- Test each gene individually using all 100 samples and identify genes with univariate p-value less than 0.001
- For $i \in 1, 2, \ldots, 100$
    - leave out the i^{th} sample
    - build LDA classifier on remaining 99 samples with selected genes
    - evaluate classifier in the i^{th} sample
- Average the LOOCV misclassification estimates across all folds

Cross-Validation

Leave one out cross validation (LOOCV) with gene selection

- For $i \in 1, 2, \ldots, 100$:
  - leave out the i^{th} sample
  - Test each gene individually using 99 samples and identify genes with univariate p-value less than 0.001
  - build LDA classifier on remaining 99 samples with selected genes
  - evaluate classifier in the i^{th} sample
- Average the LOOCV misclassification estimates across all folds

Cross-Validation



Summary

- The development and validation of predictors can be integrated into clinical trials
- Cross-validation can improve the efficiency of estimating predictors
- All data steps need to be incorporated into the validation process to avoid over estimation

Thanks

MAYO CLINIC