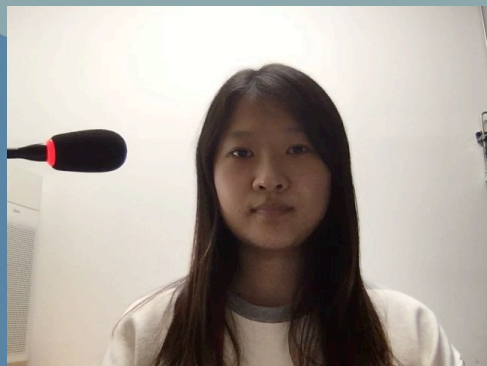
Week 10 - Predictive modeling

L10-03. Paper review – Poldrack et al. 2019

Soo Ahn Lee
Ph.D. student

30 April 2021



Poldrack et al. (2019) JAMA Psychiatry

Clinical Review & Education

JAMA Psychiatry | Review

Establishment of Best Practices for Evidence for Prediction
A Review

Russell A. Poldrack, PhD; Grace Huckins, MSc; Gael Varoquaux, PhD

IMPORTANCE Great interest exists in identifying methods to predict neuropsychiatric disease states and treatment outcomes from high-dimensional data, including neuroimaging and genomics data. The goal of this review is to highlight several potential problems that can arise in studies that aim to establish prediction.

OBSERVATIONS A number of neuroimaging studies have claimed to establish prediction while establishing only correlation, which is an inappropriate use of the statistical meaning of prediction. Statistical associations do not necessarily imply the ability to make predictions in a generalized manner; establishing evidence for prediction thus requires testing of the model on data separate from those used to estimate the model's parameters. This article discusses various measures of predictive performance and the limitations of some commonly used measures, with a focus on the importance of using multiple measures when assessing performance. For classification, the area under the receiver operating characteristic curve is an appropriate measure; for regression analysis, correlation should be avoided, and median absolute error is preferred.

CONCLUSIONS AND RELEVANCE To ensure accurate estimates of predictive validity, the recommended best practices for predictive modeling include the following: (1) in-sample

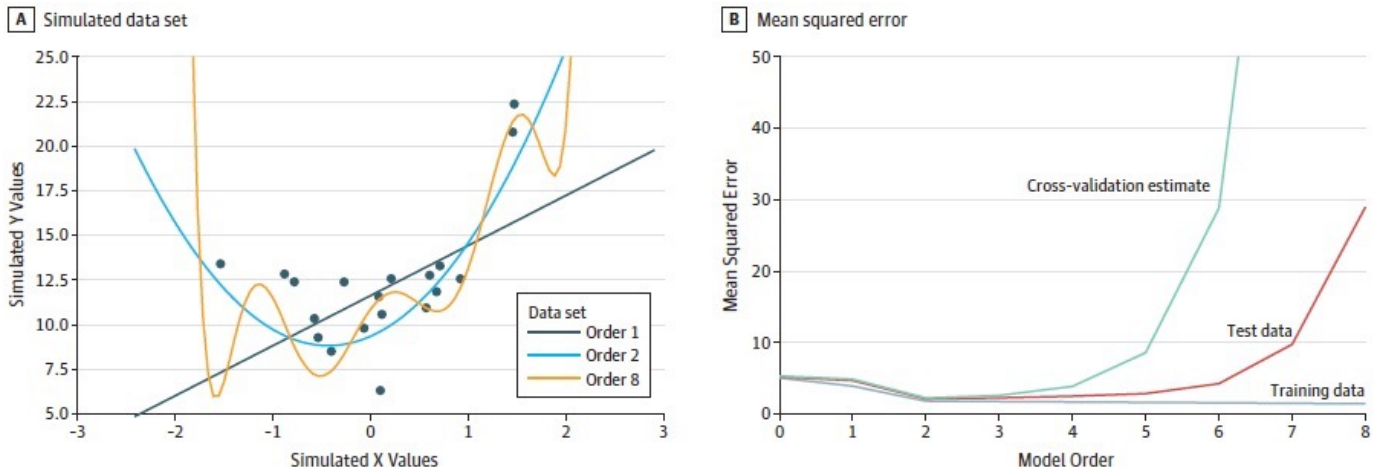
Supplemental content

Association vs. Prediction

Difference between association and prediction: prediction involves generalizability to new data

Models usually fit to the sample data than new data → overfitting

Solution: cross-validation (leave-one-out, k-fold, etc.) → uses subsets of the data to iteratively train and test the predictive performance of the model



Factors that can bias assessment of prediction

1. Small samples: smaller samples, higher variability → exaggerated predictive accuracy
2. Leakage of test data: violation of an isolation of the test data during the fitting of models to the training data
3. Model selection outside of cross-validation: necessity of a held-out validation set
4. Nonindependence between training and testing sets: data from families (e.g., HCP data), time series data (autocorrelation)

Quantification of predictive accuracy

1. Classification accuracy:
   - Use ROC curve to avoid problem from imbalance between the frequencies of different classes
   - Present sensitivity and specificity separately → can see the balance of false positives and false negatives

2. Regression accuracy:
   - Use explained variance or absolute error measurements
   - Prediction-outcome correlation → report the explained variance computed by the sum of squares together
   - Leave-one-out cross-validation → k-fold or random-split cross-validation

Best practices for predictive modeling

1. For regression analyses, measure of variance such as R^2 should be accompanied by measures of unsigned error (e.g., mean squared error or mean absolute error)
2. For classification analyses, accuracy should be reported separately for each class, and a measure of accuracy insensitive to relative class frequencies (e.g., AUC) should be reported
3. The correlation coefficient should be computed by using the sums-of-squares formulation rather than simply by squaring it
4. k-fold cross-validation (k: 5~10) can be used rather than leave-one-out cross-validation when the testing set in leave-one-out cross-validation is not representative of the whole data or anticorrelated with the training set