Predictor development and evaluation

University of BRISTOL
BBSRC bioscience for the future
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
MRC Integrative Epidemiology Unit

Workshop outline

- Introduction
- Omic data – genetic, epigenetic and metabolite
- **Prediction – methodology**
- Prediction – demonstration
- Examples and case studies

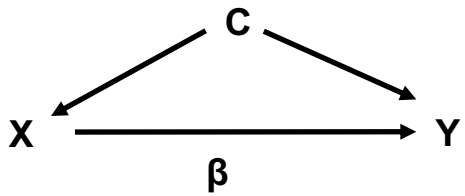
By the end of the session you will learn...

1. Aim of modeling for causation vs. prediction
2. Study design approaches for optimal prediction
3. Performance metrics for continuous, binary/categorical outcomes
4. Common pitfalls
5. Pros & cons of machine learning methods:
   - Regression-based models
   - Black box models

Aims of causal effect estimation

Researching *causal effects* and *prediction* both utilize statistical estimation

In the causal setting, we want to know how some exposure (x) causes some outcome (Y) independent of other factors (C)

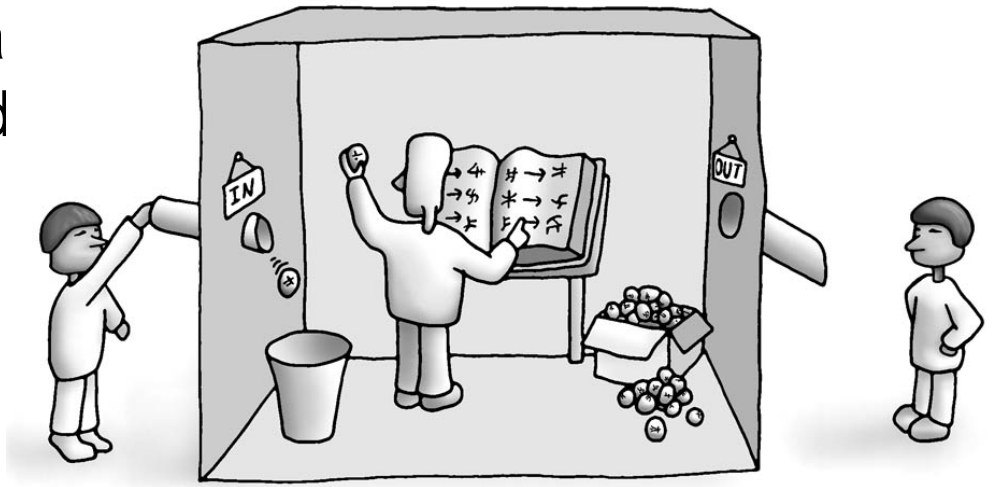


- E.g. does cigarette smoking cause a change in DNA methylation at cg05575921 in the *AHRR* gene?

- Maximize power to estimate β by using total N
- Use formal hypothesis testing or causal inference frameworks to draw conclusions

Aims of prediction

- Interested in generating a function to input observed variables (X, C) and predicting a value for Y



https://kashifwashere.com/2017/04/04/syntax-and-semantics-in-searles-chinese-room-argument/

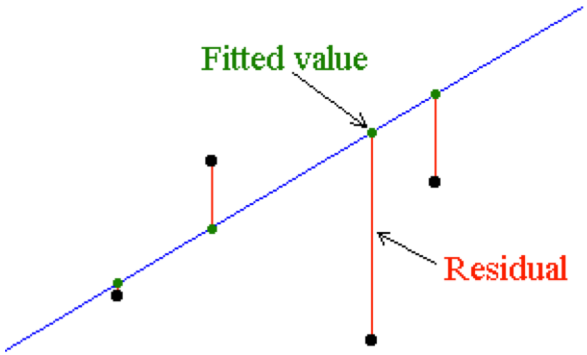
- Relationship need not be causal!
- "Best" function predicting Y is an open question
- Maximize *out-of-sample* or *generalization* performance
  - How well it predicts Y in new data when it doesn't have the answer

Prediction error

- Model fitting seeks to minimize error
  - E.g. ordinary least squares (OLS) regression minimizes the residual sum of squared (RSS) error:

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

  - Called a 'loss function'



- Model fitting for prediction seeks to minimize *generalization* or *test* error
  - Possible to measure in a single sample (AIC, BIC)
  - Most simply measured in an additional dataset

Bias, variance and model complexity

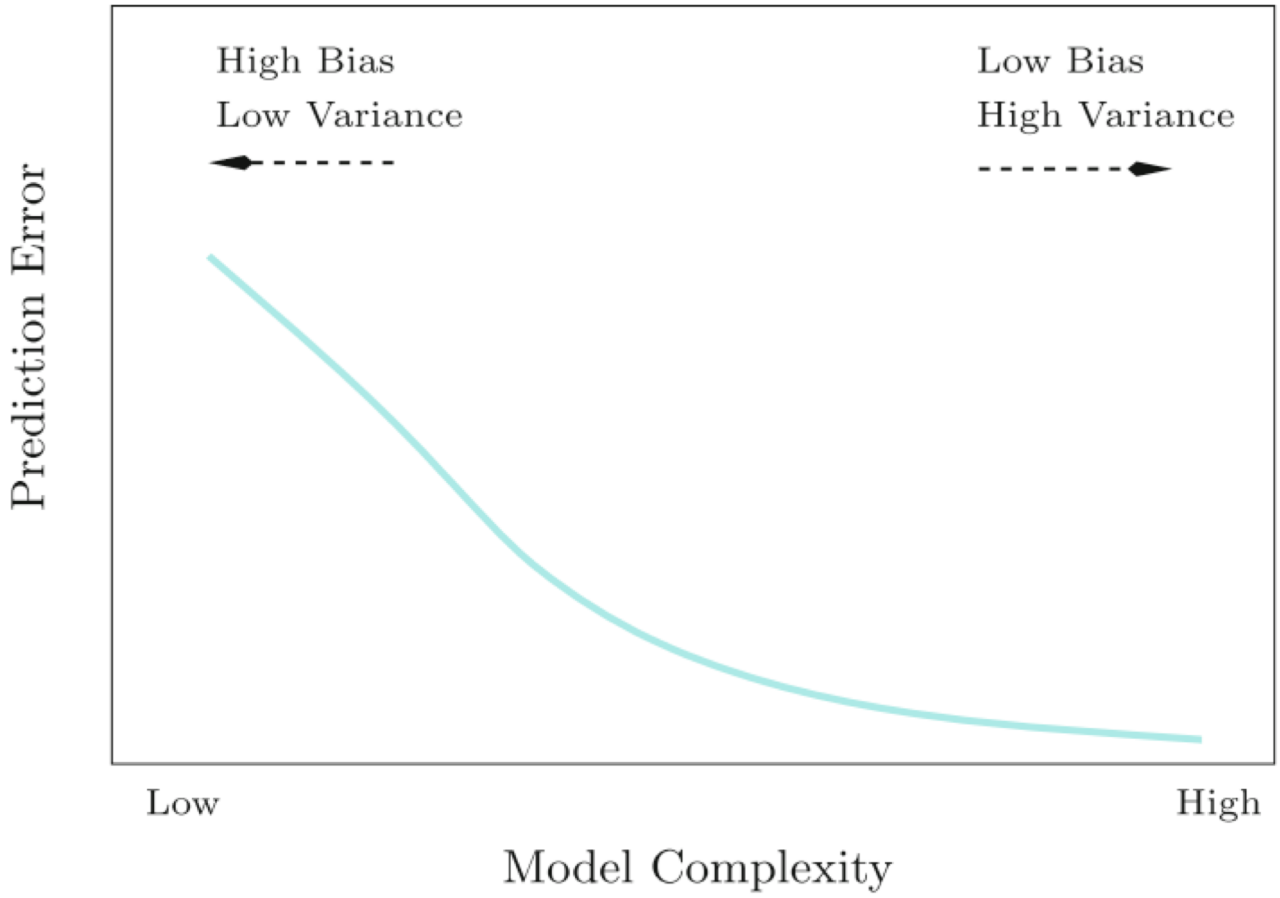
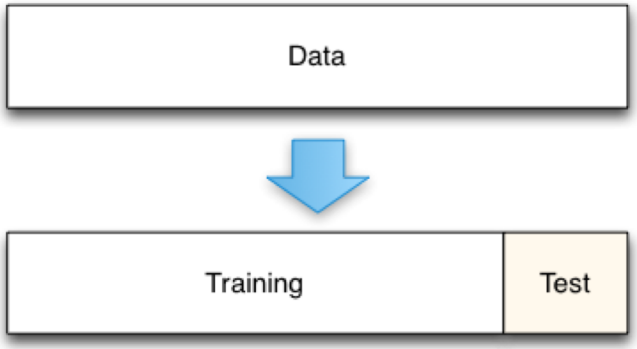


FIGURE 2.11. *Test and training error as a function of model complexity.*

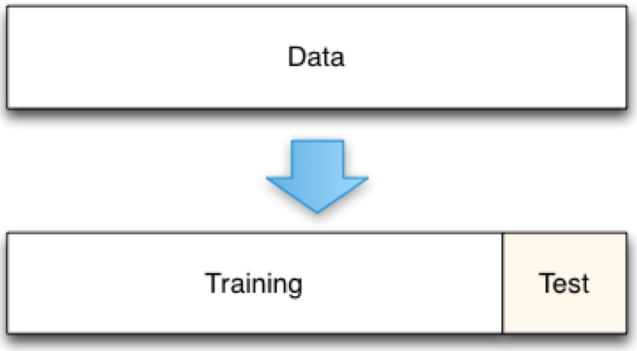
Cross validation

- The most common way to assess the out of sample prediction performance is to partition our full dataset in two sub-sets:
  - Training set
    - Develop potential models with the benefit of access to the observed true outcome
  - Testing or validation set
    - Perform each model's prediction withholding information of the true outcome
    - We can then compare the true and predicted outcomes to get a real measure of *test error*



http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2015/06/07_cross_validation_diagram.pn

Cross validation



Calculate loss function:

- *Continuous*

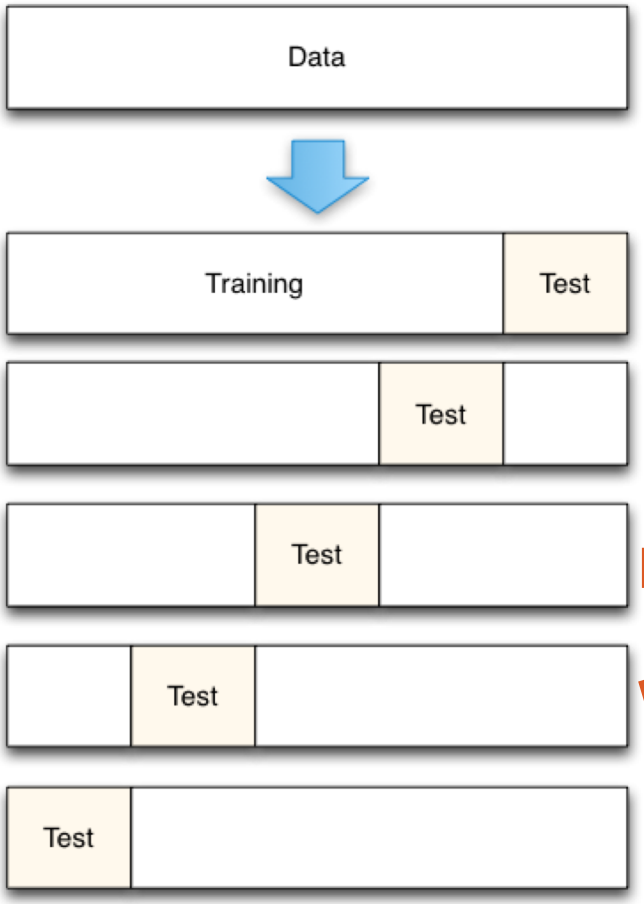
$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{squared error} \\ |Y - \hat{f}(X)| & \text{absolute error.} \end{cases}$$

- *Binary*

$$L(G, \hat{G}(X)) = I(G \neq \hat{G}(X)) \quad \text{(0–1 loss)}$$

http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2015/06/07_cross_validation_diagram.png

K-fold cross validation



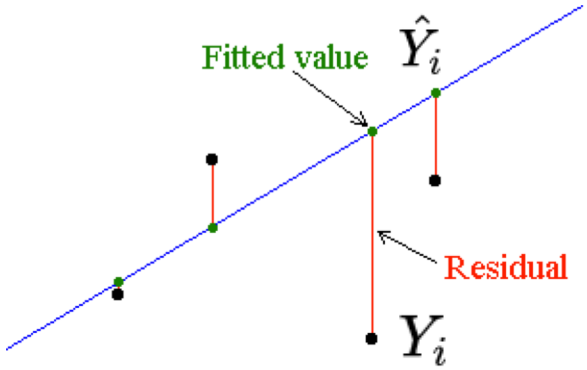
- Doing this multiple times across different chunks, or k-folds, of our the data allows a more robust assessment

Average our loss function across folds to see how we do over all

http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2015/06/07_cross_validation_diagram.png

Performance assessment, continuous outcomes

- We are essentially comparing the similarity of two vectors of numbers:
  - the true Y_i observed and the $\hat{Y}_i$ we predicted
- Mean squared error (MSE) or root-MSE
  - Root-MSE summarizes the performance on the scale of the variable we're trying to predict (Y)
- Correlation or R^2
  - Gives a summary on a familiar, standardized scale (0 to 1)
  - Is interpretable as the percent of the variance in the outcome that the predictor explains



$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Performance assessment, binary outcomes

Confusion matrix

		Reference / True Classes	
		Event	No Event
Predicted classes	Event	***A*** True Positives	***B*** False Positives
	No Event	***C*** False Negatives	***D*** True Negatives

Accuracy	(A+D)/(A+B+C+D)	% of predicted ‘event’ & ‘no event’ that are true
Sensitivity	A/(A+C)	When it is a true ”event" how often it predicts ”event"
Specificity	D/(B+D)	When it is a true ”no event" how often it predicts ”no event"

Performance assessment: ROC curves

Inst#	True class	Score	Predicted class, threshold = 0.7
1	p	.9	Y
2	p	.8	Y
3	n	.7	Y
4	p	.6	N
5	p	.55	N
6	p	.54	N
7	n	.53	N
8	n	.52	N
9	p	.51	N
10	n	.505	N
11	p	.4	N
12	n	.39	N
13	p	.38	N
14	n	.37	N
15	n	.36	N
16	n	.35	N
17	p	.34	N
18	n	.33	N
19	p	.30	N
20	n	.1	N

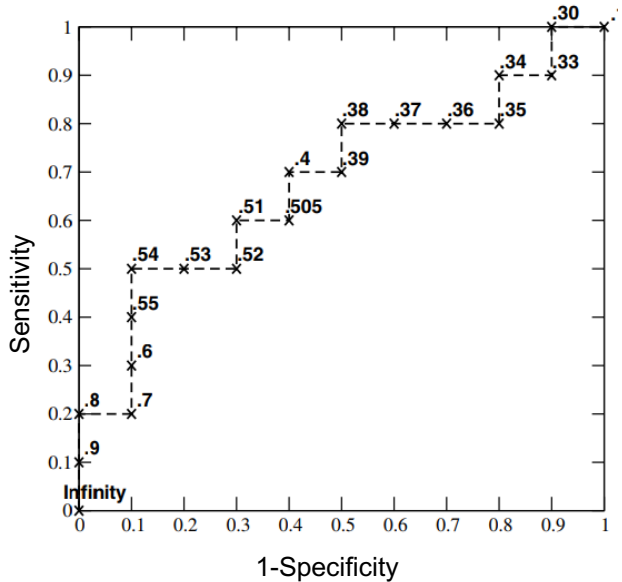
Receiver operating characteristic (ROC) curves plot how well different cut-off values of a continuous prediction score do at separating a binary outcome

- Y-axis: Sensitivity, X-axis: 1-specificity

		True Class p	True Class n
Predicted class, threshold = 0.7	Y	2	1
	N	8	9
		10	10

Sensitivity = A/(A+C)
= 2/10
= 0.2

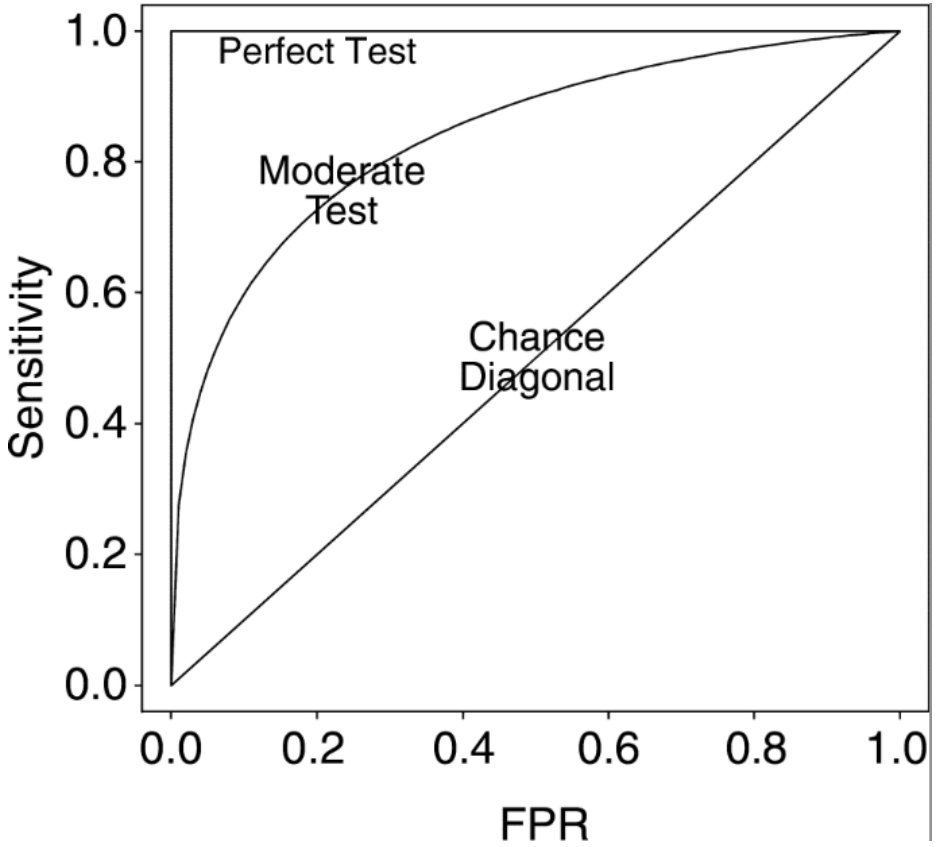
1-specificity = 1- (D/(B+D))
= 1 - 9/10
= 0.1



Fawcett, T. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874 (2006).

Performance assessment, binary outcomes

- The more the curve is toward the top-left the better were doing at separating the two classes
- Can be summarized with a single number:
  - the area under the curve (AUC), an aggregate measure of performance across the range of trade offs between false positives and false negatives



Common challenges, limitations & pitfalls

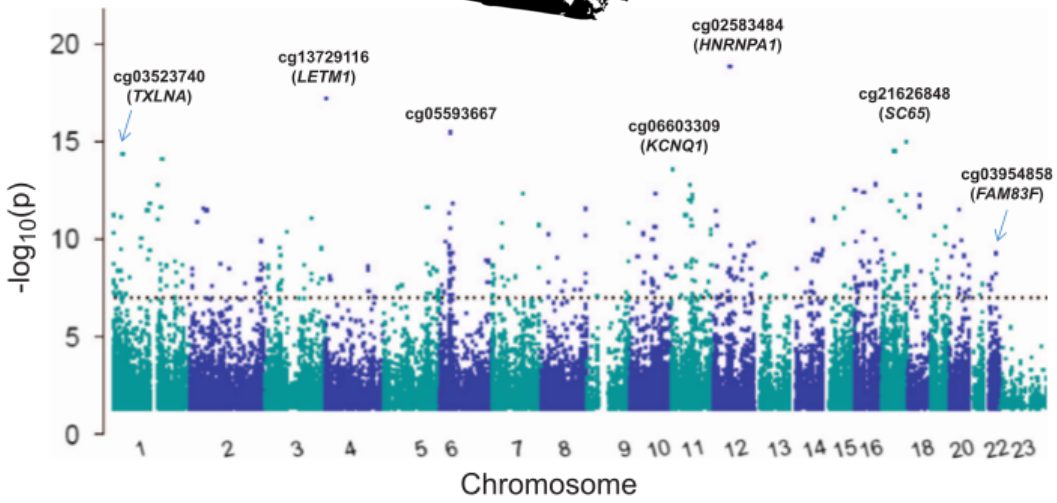
1. *Evaluate performance in independent sample*
   - Training and testing on the same observations gives inflated performance
     - E.g. DNAm Alcohol score by Liu et al. 2016
2. *Fit final model with maximum power*
   - Cross-validation should be used for model selection, but final models should be run on all data
3. *Match training and applied dataset characteristics*
   - If you train on child subjects, your predictor may not be valid for application in the elderly
     - E.g. gestational age DNA methylation clocks by Knight et al. 2016 & Bohlin et al. 2016

DNAm alcohol score

COHORTS FOR HEART AND AGING RESEARCH IN GENOMIC EPIDEMIOLOGY

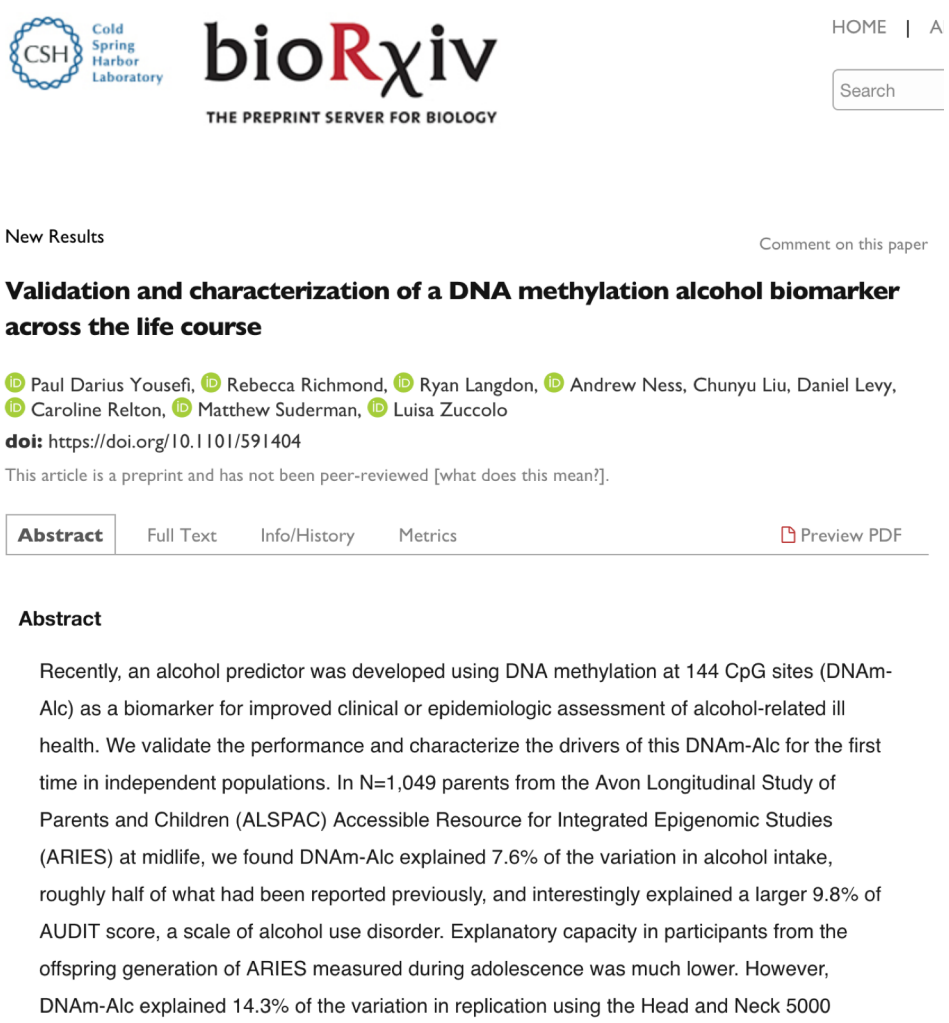
Liu, C. *et al*. *Mol. Psychiatry* (2016)

- Meta-analysis of peripheral blood DNA methylation
- 13,317 individuals in13 cohorts
- 328 CpG sites in European ancestry at $p < 1x10^{-7}$



- Made an alcohol intake score using 144 CpGs from a lasso model on top CpGs in preliminary analysis
  - R^2 = 12.0 – 13.8% of alcohol intake
  - AUC=0.90 – 0.99 for heavy vs. non-drinkers

DNAm alcohol score



Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | A

Search

New Results

Comment on this paper

Validation and characterization of a DNA methylation alcohol biomarker across the life course

Paul Darius Yousefi, Rebecca Richmond, Ryan Langdon, Andrew Ness, Chunyu Liu, Daniel Levy, Caroline Relton, Matthew Suderman, Luisa Zuccolo

doi: https://doi.org/10.1101/591404

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | Full Text | Info/History | Metrics | Preview PDF

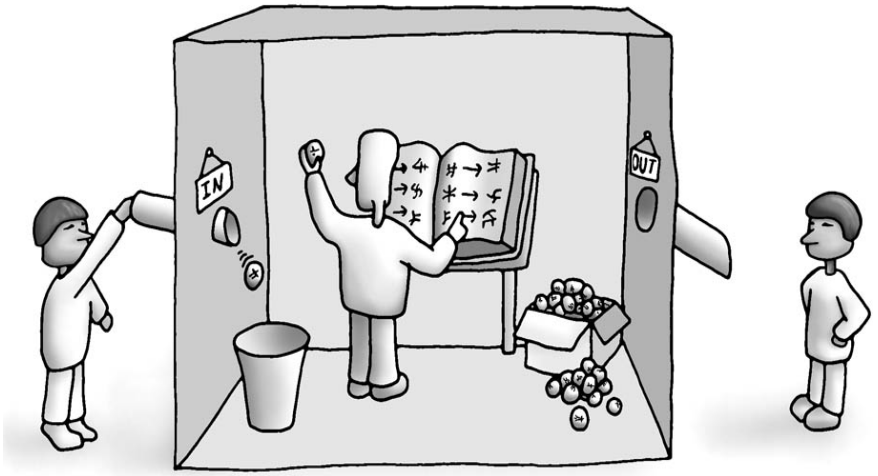
Abstract

Recently, an alcohol predictor was developed using DNA methylation at 144 CpG sites (DNAm-Alc) as a biomarker for improved clinical or epidemiologic assessment of alcohol-related ill health. We validate the performance and characterize the drivers of this DNAm-Alc for the first time in independent populations. In N=1,049 parents from the Avon Longitudinal Study of Parents and Children (ALSPAC) Accessible Resource for Integrated Epigenomic Studies (ARIES) at midlife, we found DNAm-Alc explained 7.6% of the variation in alcohol intake, roughly half of what had been reported previously, and interestingly explained a larger 9.8% of AUDIT score, a scale of alcohol use disorder. Explanatory capacity in participants from the offspring generation of ARIES measured during adolescence was much lower. However, DNAm-Alc explained 14.3% of the variation in replication using the Head and Neck 5000

- Only selected their CpGs using the lasso model in their training set
- Re-estimated the actual β values in the testing set
- Artificial inflation of prediction performance due to over-fitting in test data
- We show performance roughly half of initial report (R^2 = 7.6) for intake, AUC = 0.6

Interpretable prediction models

- Often we do care about what features drive our predictions
  - Does model make biological sense?
- Trivial in the case of multiple regression
  - β's summarize variable importance

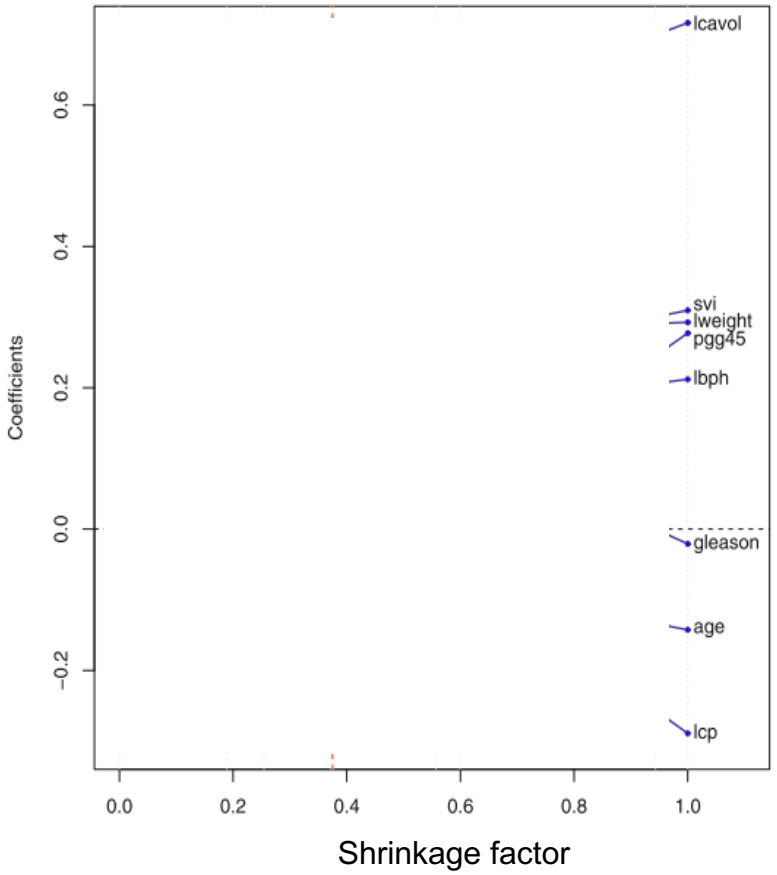


https://kashifwashere.com/2017/04/04/syntax-and-semantics-in-searles-chinese-room-argument/

- Regression models can include too many, often redundant, variables that give low bias/high variance predictions
  - Especially in omic contexts with thousands of inputs

Shrinkage regression models

- Estimates interpretable β coefficients like ordinary least squares (OLS) regression
  - Major difference: keeps on the β's of the most important predictors
  - Less important β's are 'shrunk' to 0 and dropped from the model
- Application to continuous, classification and survival problems
- Examples include:
  - *Lasso, ridge, elastic net, LAR, principal component & partial least squares regression*



Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition (Springer Science & Business Media, 2009).

Penalized regression

- Lasso, ridge, and elastic net perform β shrinkage by penalizing their total magnitude with parameter λ:

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

- Optimal λ selected through cross-validation
- For lasso: q = 1, for ridge: q= 2
- Elastic net balances both by adding mixture parameter α:

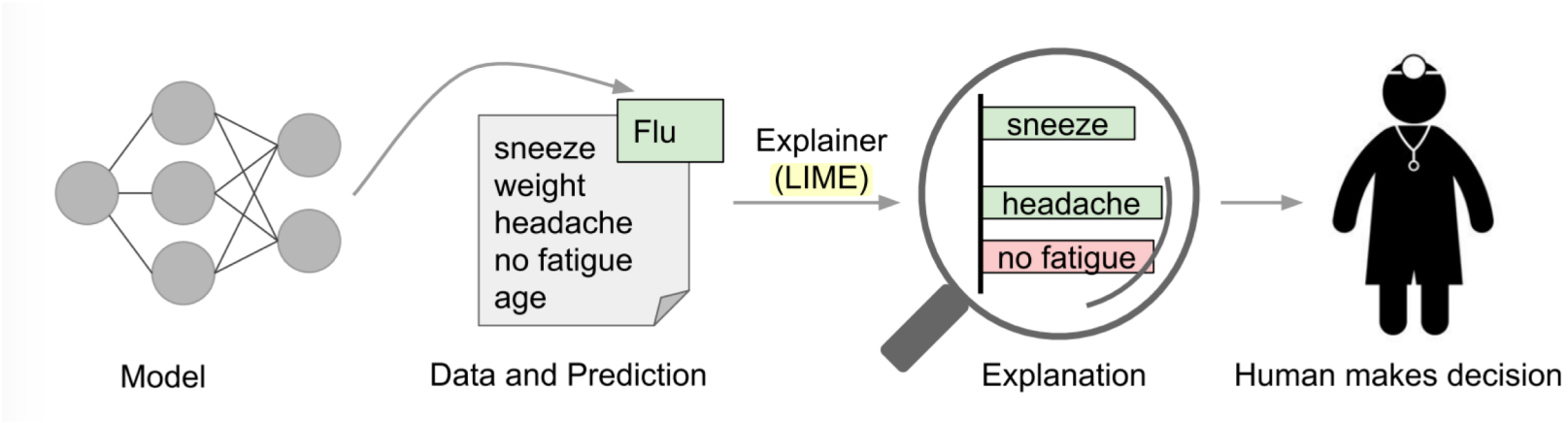
$$\lambda \sum_{j=1}^{p} \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j| \right)$$

Black box models

- Many supervised and unsupervised approaches to choose
  - growing by the day!
- Valid for omic applications
- Performance assessment by cross-validation
- Parameter optimization by cross-validation and/or resampling
- Major issue is interpretation

Nearest Neighbors	• K-nearest neighbors, nearest centroid, neighborhood component analysis
Support vector machines	• linear SVM, Kernal SVM
Neural networks	• Bayesian regularized, model averaged, multi-layer perceptron
Ensemble methods	• Random Forests, Gradient Boosting, AdaBoost, SuperLearner

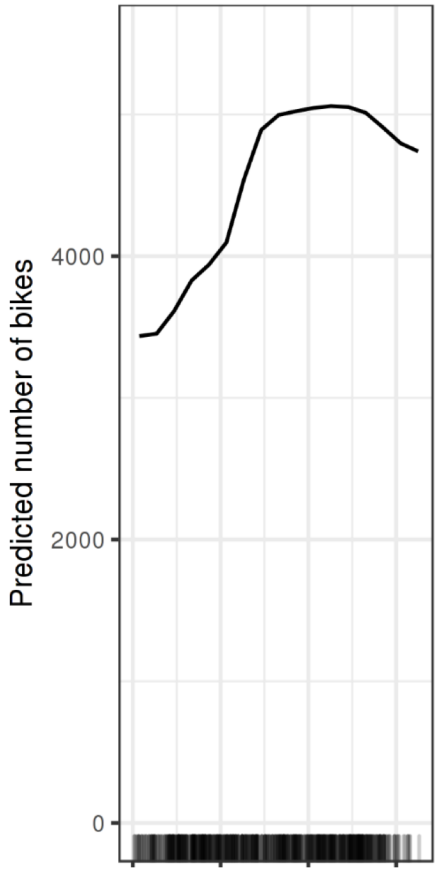
Opening the black box

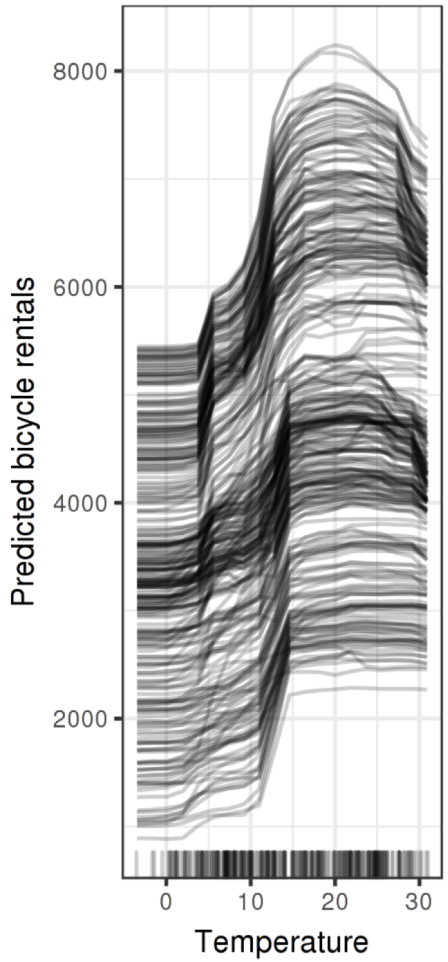


Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv* 2016. http://arxiv.org/abs/1602.04938

- Increasing emphasis on post-hoc interpretation
- Methods development expanding:
  - Partial dependence plots (PDP)
  - Individual conditional expectation (ICE)
  - Accumulated local effect plots (ALE)
  - Local Interpretable Model-agnostic Explanations (LIME)

Opening the black box





$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_S, x_C^{(i)}\right)$$

Molnar, C. Interpretable Machine learning: A Guide for Making Black Box Models Explainable. 2019. https://christophm.github.io/interpretable-ml-book

Iml r-package https://cran.r-project.org/web/packages/iml/

Conclusions

1. Aim of modeling for causation vs. prediction
   - *In-sample vs. prediction error*
2. Study design approaches for optimal prediction
   - *Cross-validation*
3. Performance metrics for...
   - Continuous outcomes: *MSE/Root-MSE, R^2*
   - Binary outcomes: *Balanced accuracy, no-information rate, AUC*
4. Common pitfalls
5. Pros & cons of regression-based vs. black box machine learning methods
   - *Post-hoc interpretability*
   - *Limited complexity and reliance on user knowledge*

Workshop outline

- Introduction
- Omic data – genetic, epigenetic and metabolite
- Prediction – methodology
- **Prediction – demonstration**
- Examples and case studies