Introduction to Machine Learning in Epidemiology

Andrea Bellavia, PhD

TIMI Study Group, Brigham and Women's Hospital, Harvard Medical School
Department of Environmental Health, Harvard T.H. Chan School of Public Health

abellavia@bwh.harvard.edu

Karolinska Institutet

September 2025

Table of contents

1. Introduction

1.1. Complex data in modern epidemiology

- Big data revolution drastically affecting epidemiology and clinical research
- Large amount of data that can be rapidly collected and can help identifying complex patterns and make more informed decision
- Example of big data that can impact clinical research:
  - Omics data
  - Hemodynamic parameters
  - Imaging
  - Data from wearable devices
  - ...
- Complex data does not only refer to the number of evaluated factors but also to the biological mechanisms involved (e.g. non-linear effects)

New research paradigms

Novel frameworks originating from the big data revolution, such as:

- Precision medicine: "Precision medicine, sometimes known as personalized medicine, is an innovative approach to tailoring disease prevention and treatment that takes into account differences in people's genes, environments, and lifestyles."
- Exposome: "the totality of environmental exposures, both internal and external, that a person experiences throughout their life, and how these exposures impact their health"[1]

[1] See Vermeulen et al, Science 2020

Analytical challenges

- To what extent can our classical statistical approaches be used to accommodate the complexity of these novel scientific frameworks?
- Growing interest in analytical approaches that can handle this complexity while mitigating overfitting
  - Semi-parametric and non-parametric regression
  - Machine Learning
  - Deep learning
- Specific challenges when applying these methods to the analysis of epi data

Complex data and clinical prediction

- Integrating big data in clinical prediction poses important issues:
  - How can we blend the interest in finer individual precision with the need of a parsimonious and pragmatic model?
  - Are there settings where one might favor one vs the other?
  - Does data complexity always come with complex interpretability?
  - How do I present complex results to peers? Or to patients?
- Moreover, as one starts incorporating omics data into individualized prediction, complex features such as biological interactions might need to be accounted for. Can we do that with regression?

Complex data and etiological research

- Similar challenges arise with etiological research (e.g. in the exposome framework):
  - How can we maintain the causal inference rigor required by etiological studies when incorporating complex predictors?
  - How do we account for confounding, effect modification, or even mediation?
  - How do we treat social determinants of health?

Motivating example 1 [clinical prediction]

- Research grant on proteomic discovery for heart failure risk in diabetic patients[2]
- $\sim$ 70 cases and 70 controls
- Blood sample used to quantify $\sim$ 400 proteins
- Validation of selected proteins in a larger cohort of 14,000 trial participants

Analytical challenge: how to identify the top predictors (out of $\sim$ 400) with only 70 events?

- Individual regressions are unreliable due to co-confounding and require multiple comparison assessment
- Multivariable models will not converge ($p > n$)
- Risk of overfitting

[2] Main results here

Motivating example 2 [exposomic research]

- Identification of potential risk factors for ALS[3]
- Interest in evaluating history of medication
- 501 ALS cases and matched controls
- Data on over 1000 medications

Analytical challenge: can we use these data to identify potential risk factors?

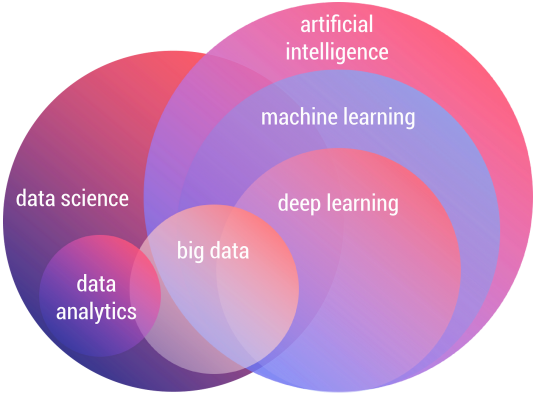
- In addition to the items discussed in the previous slide we should also consider confounding, causal interpretation, and potential translation to intervention (etiological study)

[3] Rotem et al 2024

Goals of this course

- Understand the extent to which classical analytical tools can be used to address the nature of complex epidemiologic data
- Introduction to general topics and methods in machine learning
- Specific challenges of ML application in epidemiology and how to address them

1.2. What is Machine Learning?



General introduction, link

1950's definition: *the field of study that gives computers the ability to learn without explicitly being programmed*

- Terminology (machine learning, deep learning, data science, statistics . . .) can be confusing.
- Clinical researchers and epidemiologists tend to think of ML as an alternative to classical regression in a black and white fashion
- This is a broad and incomplete generalization, which can provide misleading messages and unnecessary skepticism
- The separation line is not defined by the approaches themselves but rather in terms of data assumptions. From Breiman 2001, "The two cultures":
  - One [culture] assumes that the data are generated by a given stochastic data model
  - The other [culture] uses algorithmic models and treats the data mechanism as unknown

The regression vs ML debate

- Why do we tend to think of regression and ML as 2 different worlds?
- This separation goes back to a seminal paper (2001) by Leo Breiman titled "Statistical Modeling: The Two Cultures"
- The separation line, however, is not defined by the approaches themselves but rather in terms of data assumptions. From the abstract:
  - One [culture] assumes that the data are generated by a given stochastic data model
  - The other [culture] uses algorithmic models and treats the data mechanism as unknown

1.3 Bridging statistics and ML

- Using Brieman's terminology, a classical regression model makes strong assumptions about the underlying stochastic data model (e.g. normality, linearity, additivity)
- Inference on parameters such as β coefficients, OR, HR, is valid under these assumptions
- There are situations where we might want to relax at least some of these assumption
- To relax assumptions, however, we do not need to necessarily change the modeling framework
- Next: discuss the extent to which classical regression can handle complex data in epi research

2. Classical Regression, complex data, and model overfitting

2.1 Linearity and additivity

- Regression modeling is ubiquitous in clinical and epidemiological research to connect one or more covariates (ind. variables) and a certain health outcome (dep.)
- Basic idea: identify a functional form between dependent and independent variables
- Commonly, these functional forms involve linear relationships

Common examples

- Linear regression [linear on the expected value of a continuous outcome]

$$E[Y] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

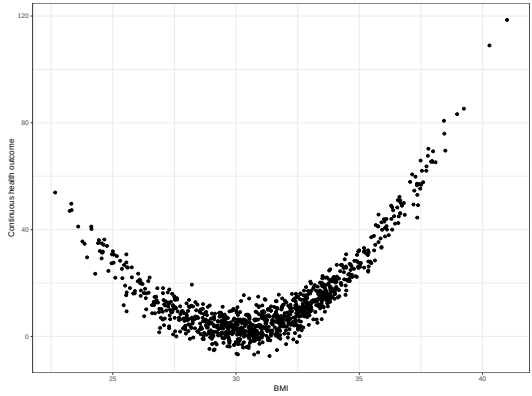
- Logistic regression [linear on the logarithm of the odds (logit) of a binary outcome]

$$logit(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Cox regression [linear on the logarithm of hazard ratio for a time-to-event outcome]

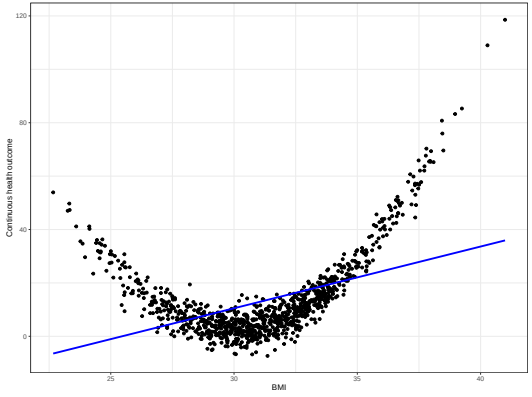
$$log(HR) = \beta_1 X_1 + \cdots + \beta_p X_p$$

- The use of linear functions implies that specific assumptions are made for continuous predictors
- Example: continuous predictor (e.g. BMI) and continuous outcome Y



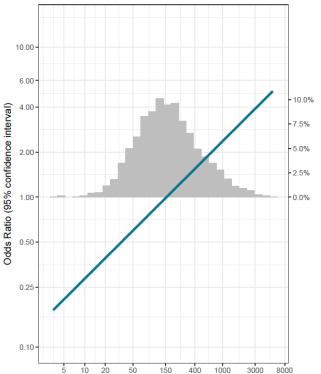
Linearity assumption in practice

- Fit a linear regression: $E[Y|\text{BMI}] = \beta_0 + \beta_1 \cdot \text{BMI}$. Result: $\hat{\beta}_1 = 2.3$
- Interpretation: difference in $E[Y]$ **for each** unit increase in BMI [blue line]
- The effect is the same when we compare BMI of 21 vs 20, 11 vs 10, 56 vs 55 etc.



Log-linear models (e.g. logistic, Cox)

- Linearity assumptions have slightly different interpretations in logistic and Cox model, which define linear assumptions on the logarithmic scale (log-linear models)
- Plotting figures *on the log-scale* is required to visualize linearity (and, later, potential departures)



Additivity

- Regression models make several additional assumptions (e.g. residuals normality, homoscedasticity, proportionality of the hazards ...)
- Additivity is another silent assumption with relevant implications on results' interpretation and translation
- Additivity assumptions are made for any combination of covariates included in a regression model

Additivity: implications

The assumption of additivity between two covariates implies:

- Their joint effect equals the sum of the two main effects: absence of interaction
- The effects of each covariate are constant over levels of the other covariate: absence of effect modification

Note: in log-linear models, additivity assumption translates into a multiplicative assumption on the OR and HR scale [4]

[4] For more details see: VanderWeele TJ, Knol MJ. A tutorial on interaction. Epidemiologic methods. 2014 Dec 1;3(1):33-72., / Bellavia and Murphy. Clinical interpretation of statistical interaction. Circulation, 2025

Relaxing additivity

This is more straightforward: inclusion of a product term (aka interaction term) relaxes assumption of additivity and can be used to assess interaction or effect modification

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2$$

- β_3 can be used to test for interaction and/or effect modification
- Interpretation of the results should take into account the scale (i.e. additive interaction vs multiplicative interaction)

2.2 Relaxing linearity using Splines

A common approach to relax the linearity assumption is by creating a categorical version of the continuous covariate, included in regression models using dummy variables

Example: create 4 groups using quartiles of the distribution

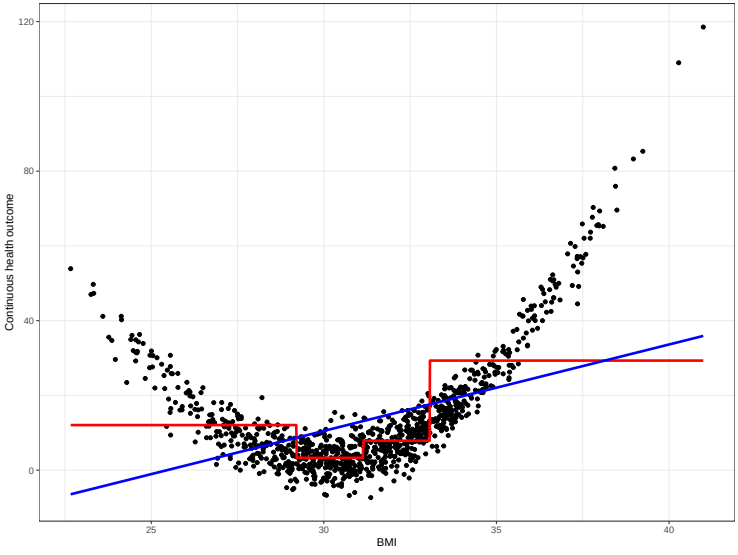
- Old (linear) model: $E[Y] = \beta_0 + \beta_1 X_1$
- New (categorical) model: $E[Y] = \beta_0 + \beta_1 X_{25th-50th} + \beta_2 X_{50th-75th} + \beta_3 X_{75th-100th}$

with $X_{25th-50th}, X_{50th-75th}$, and $X_{75th-100th} = (0,1)$

Table 1: Continuous and categorical version

x1	x1cat
2.7565865	3
3.1283981	4
0.6720901	2
-0.3604915	2
2.8176624	3
-4.4538679	1

Categorization in practice



We are replacing an assumption (linearity) with another assumption (step function)

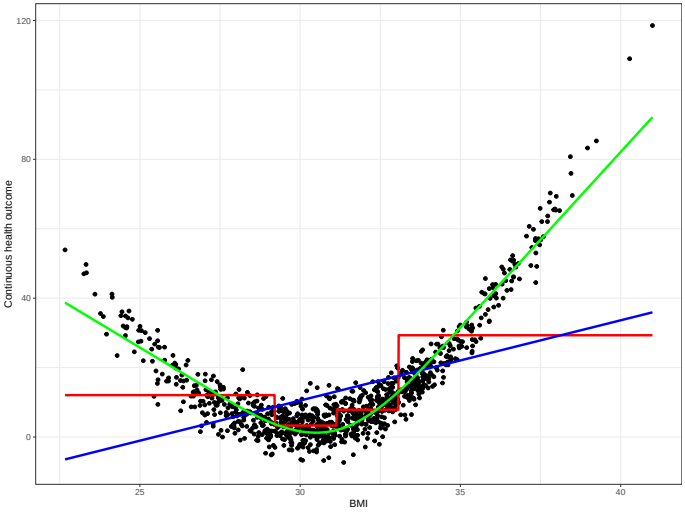
- We now assume that the predicted response will be exactly the same for all individuals in the same subgroups
- We are also assuming that the change in the outcome will occur at specified (a priori and often subjectively) jumps
- Using the two cultures definition, we are still grounded within a stochastic model for the data based on clearly defined assumptions

Issues with categorization have long been recognized

- Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 1995
- Greenland S. Problems in the Average-Risk Interpretation of Categorical Dose-Response Analyses. Epidemiology 1995.
- Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006

Splines

A more flexible solution: splines modeling



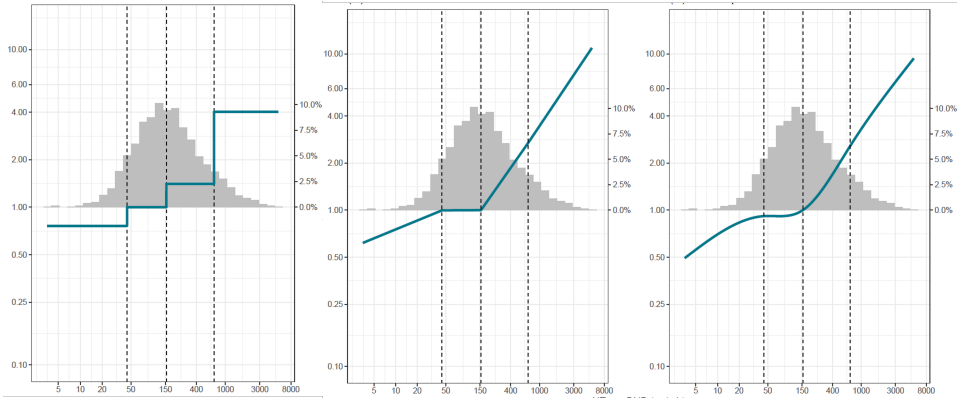
Splines transformations involve 2 steps:

- Select how many knots and where to place them
  - Conventionally at distribution percentiles. In general, 3 or 4 might suffice. Explore more with skewed distributions or if there is specific interest at the tails
- Select how to model in between knots

This interactive website provides a great tool to understand more the different assumptions and impact of knots numbers and locations: link

How to model in between knots?

- Categorical analysis is actually a particular case of splines modeling (degree 0 splines) where we assume constant outcome levels between knots
- Alternatively, we could use linear function that changes slope at each knot (degree 1 splines, aka piecewise modeling)
- Degree 3 (cubic) splines: between knots, the curve is a cubic polynomial, a smooth function of the form $\beta_0 + \beta_1 x + \beta_2 x + \beta_3 x$



Degree 0 (left, categorical approach), degree 1 (center, piecewise linear model, and degree 3 (right, cubic) splines for modeling the association between a continuous predictor and a binary outcome (OR scale)

Restricted Cubic Splines

- Restricted cubic splines add a constraint of linearity before the first and after the last knot
- Practically, the continuous covariate is replaced by $k-1$ new variables, where k is the number of knots. A key feature of RCS is that the first of these new variables coincide with the original covariate:

$$s_1 = x$$

$$s_i = \frac{(x-t_{i-1})_+^3 - (x-t_{n-1})_+^3 \frac{(t_n - t_{i-1})}{(t_n - t_{n-1})} + (x-t_n)_+^3 \frac{(t_n - t_{n-1})}{(t_n - t_{n-1})}}{(t_n - t_1)^2}$$

Table 2: RCS transformation with 3 knots. The original covariate (first column) is replaced by 3-1 new variables, of which the first once coincides with the original covariate

Original predictor	1st splines transform.	2nd splines transform.
2.7565865	2.7565865	2.7126776
3.1283981	3.1283981	3.2072301
0.6720901	0.6720901	0.6792160
-0.3604915	-0.3604915	0.2281211
2.8176624	2.8176624	2.7919021
-4.4538679	-4.4538679	0.0000000

Example of R implementation and output interpretation

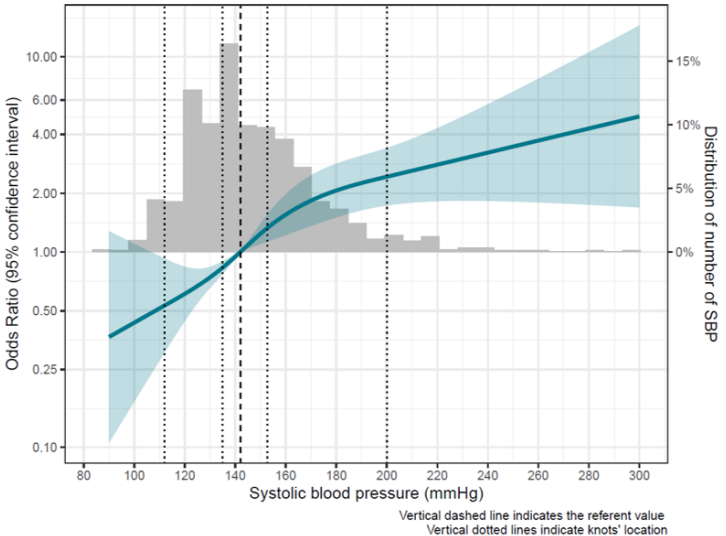
```
rcs<-glm(y~rcs(x1,3),data=cov)

round(summary(rcs)$coefficients,3)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.929      0.222  -8.694        0
## rcs(x1, 3)x1     -5.541      0.109 -50.912        0
## rcs(x1, 3)x1'    10.240      0.126  81.303        0
```

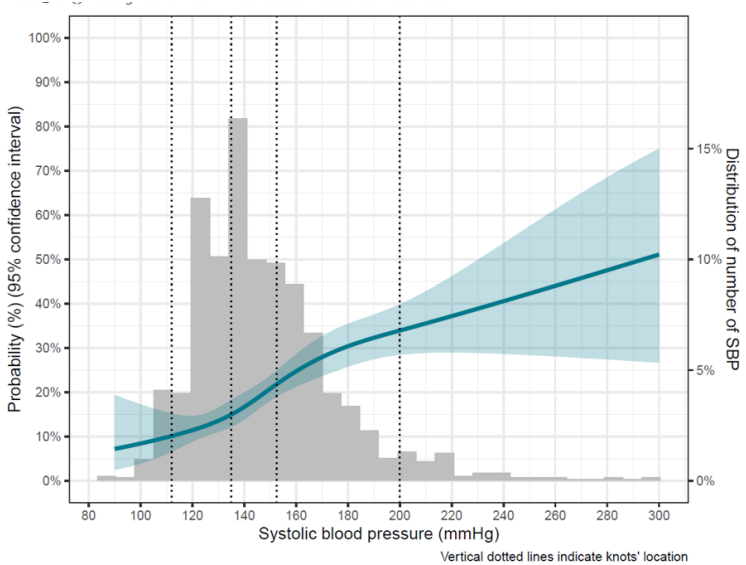
- Because of the interpretation of the first term (corresponding to the original X), we can interpret the second term as the *"non-linear term"*
  - A coefficient for this term close to 0 implies negligible departure from linearity
  - The p value can be used (with caution) as a statistical test for linearity

- The actual parameters, however, do not carry a clear interpretation.
- We need a graphical display to describe the non-linear association
- Results from the previous slide can be used to compute measures of interest and CIs across the continuous predictor. See section 5 for software material to reproduce these figures
- Example with logistic regression:



OR of CHD as a function of SBP, modeled with RCS (4 knots)

- With models such as logistic or Cox, we can either plot the OR (or HR) as a function of X, or model predictions such as the predicted probability (or absolute risk)
  - When presenting ORs, it is important to select a meaningful comparison point (e.g. the median.)
  - Note that, CIs for predictions should not cross, while CIs for ORs should cross at the reference value (OR=1)



Probability of CHD as a function of SBP, modeled with RCS (4 knots)

Recently published educational paper [Discacciati et al. 2025, IJE for restricted cubic splines in epidemiology, with code and software material available here]

JOURNAL ARTICLE

Estimating and presenting non-linear associations with restricted cubic splines Get access

Andrea Discacciati, Michael G Palazzolo, Jeong-Gun Park, Giorgio E M Melloni, Sabina A Murphy, Andrea Bellavia

International Journal of Epidemiology, Volume 54, Issue 4, August 2025, dyaf088, https://doi.org/10.1093/ije/dyaf088

Published: 17 June 2025 **Article history**

Splines and ML

- Splines transformations are a first example of bridging the two cultures between classical statistics and ML
- When operating categorization we were replacing an assumption (linear functional form) with another (step function)
- With splines, instead, we are relaxing the assumption and we are letting the machine tell us what the functional form in our data is
- Note, however, that splines modeling still requires defining which covariates should be modeled as non-linear, and the formulation of related parameters in regression models

2.3 Model overfitting

- By letting the machine tell us what happens in our data we run a severe risk: the algorithm might give an impeccable description of what is happening in our data, but the results have very poor generalizability

Model overfitting: a machine learning model performs excellently in our data but has poor performance in external data

- Suppose we have p continuous predictors in a regression model. We could include:
  - Splines transformations for each of them
  - 2-way and possibly higher order interactions between various combinations of predictors
- These would relax assumptions of linearity and additivity, but would also detect irrelevant noise and details
- The goal is to identify underlying and generalizable patterns in the data
- A robust control of overfitting is therefore key in any attempt to evaluate complex data and, as such, a cornerstone concept in ML

Additional challenges

Additional challenges that arise when attempting to use regression for overly complex data are noteworthy:

- A regression model requires a number of covariates (p) smaller than the number of individuals (n or, if the outcome is binary, the n of events). Settings where $p > n$ cannot be addressed with regression
- That is a limit: in practice, covariates/events ratios of $1/10$ or even $1/20$ are generally recommended
- Of relevance, these ratios apply to the n of parameters in the model. Including transformations and interaction terms will affect the ability of the model to converge
- Moreover, as several continuous covariates are incorporated, additional issue such as multicollinearity might arise

A note on univariate analysis for high-dimensional data

- There are settings where the data complexity is mostly driven by the exceedingly high number of potential predictors
- In genome-wide-association studies (GWAS), for example, p can easily exceed 1,000,000
- Univariate analysis (p independent regression models) is commonly used as a first screening in these settings
- P-values are adjusted for multiple comparisons (using common techniques such as Bonferroni and FDR, or permutation tests)

- We will not discuss unviariate analysis in this course because predictors in epidemiologic studies (e.g. various components of the exposome, clinical biomarkers of a certain disease) are seldom independent on each other
- Rather, the complex relationship between them (e.g. interactions) is often a target of the analysis. Moreover, we will need to incorporate confounders in the analysis
- The several predictors of interest in epi research generally need to be incorporated in a multivariable model
- When multiple regression fails to accommodate such complexity, univariate analysis with FDR adjustment is seldom a reliable choice in epidemiologic studies

2.4. Cross-validation and other methods for overfitting control

There are several ways[5] we can prevent and control overfitting in ML.

- Penalized measures (e.g. AIC, BIC) for model comparison
- Data splits and cross-validation
- Penalization and regularization
- Variable selection
- Hyperparameters tuning
- Ensemble methods

[5] Formal definition for all these to follow

AIC and BIC

- These are simple tools that can already be used with comparing regression models
- Example: should we include a splines transformation for a continuous predictor?
- Indexes like the R^2 will generally increase as reflecting better data fit
- AIC and BIC are measures of model comparison that penalize models with more parameters, thus balancing model fit and parsimony
- AIC (Akaike Information Criteria) penalizes the log-likelihood by 2 times the number of parameters: $-2 \cdot ln(L) + 2k$
- BIC (Bayesian Information Criteria) further includes n in the penalization: $-2 \cdot ln(L) + 2k \cdot ln(n)$[6]

[6]Note that, because of the "-" sign, lower values indicate better fit for both AIC and BIC

Data splits

- A first way to prevent overfitting is by splitting the data into a training set, where you build the model, and a test set where you test its performance[7]
- Common proportions are 1/2 - 1/2 or 2/3 - 1/3
- This simple data split is commonly used for the internal validation of regression-based clinical prediction models[8]

[7] When hyperparameters tuning is also included, a split into three parts is generally used. More on this later

[8] See for example Moura et al. 2023

Bootstrap resampling

Alternative to the single split, introduced by Efron in 1979 and widely regarded as one of the most revolutionary concepts in statistics[9]

- Computationally intensive technique for making inference about unknown parameters
  - Example: shape of sampling distribution for measure of association
  - Example: standard Error for measure of association
- Completely empirical
  - Based on a large number of repetitive computations

[9](https://www.amstat.org/news-listing/2021/10/08/international-prize-in-statistics-awarded-to-bradley-efron)

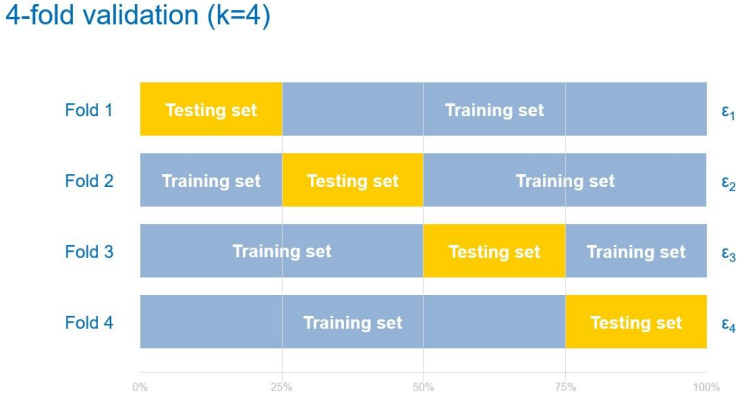
Bootstrap[10] is a method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample

- Usually at least 1000 to 5000 samples
- Each resampling has the same size of the original sample
- The parameter of interest is estimated averaging over the samples
- Replacement is allowed

[10] If interested, more details and code in 5.3.5 of Harrell's online textbook

Cross validation

CV is an extension of the single training/test split



Improvement over single training/test split and no sacrifice in sample size

Example

- Split data at random into 10 tenths
- Leave out 1/10 of data at a time
- Develop model on 9/10 and evaluate on the remaining 1/10
- Save measure of interest and repeat
- Average measure of interest over the 10 replications
- Repeat several times and average throughout

Need several repetitions (eg 50) of 10-fold cross-validation to ensure adequate precision -> computationally intensive

Hyperparameters

- Key concept for several reasons including the study of the internal mechanisms of ML and for its potential in epidemiological research
- We discussed how in ML we let the machine explore and learn from the data
- However, there are configuration settings we can control ourselves and set before the training process
- These are the so-called hyperparameters, used in the phase of model tuning
- Different ML algorithms have different sets of hyperparameters that can be defined

Hyperparameters tuning in practice:[11]

- 1. Set a grid of hyperparameters before the training phase
- 2. Split the data into two parts: training and test

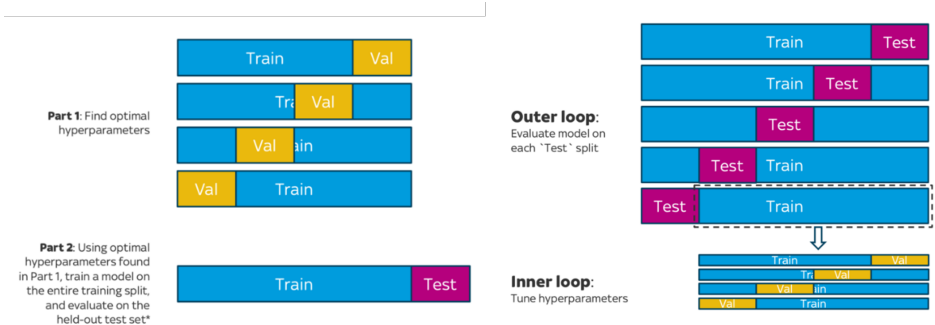
- 3. Use the training portion of data to train different ML models under the various combinations of hyperparameters and evaluate their performance in the test

[11] Figures in this section taken from this article

We could strengthen overfitting control even more by including an additional level of validation using a 3-splits system:

- 3a. Identify the best hyperparameters configuration by testing in "validation" the performance of several models trained in "train"
- 3b. Train the final model with the optimal hyperparameters in "train+val" and test its performances in "test"

Or, even better, use another *layer* of cross-validation, to prevent the particular split to affect results. CV can be incorporated in ML as a 1-layer or as a 2-layers (aka nested CV) procedure, with the latter being recommended:


Part 1: Find optimal hyperparameters
Train
Val
Val
Train
Part 2: Using optimal hyperparameters found in Part 1, train a model on the entire training split, and evaluate on the held-out test set*
Train
Test
Outer loop:
Evaluate model on each `Test` split
Test
Inner loop:
Tune hyperparameters

Summary so far

Two key considerations that have to be taken into account when assessing complex data:

1. Identifying patterns in the data might be more feasible than imposing distributional assumption
2. If this route is chosen, approaches for overfitting control and prevention are required. The more complex the setting, the more robust the approach should be (e.g. nested CV)

Supervised vs unsupervised learning

Key distinction in Machine Learning: two ways we can learn these patterns from the data

- Supervised Learning
  - Data includes outcome of interest
  - Goal is usually to predict/explain that outcome
  - Develop and validate an algorithm to predict outcome based on a set of potential predictors
  - Example outside of epidemiology: sports betting
- Unsupervised Learning
  - No outcome
  - Goal is to describe associations and patterns in high-dimensional data
  - Example outside of epidemiology: online recommendations (Youtube, Instagram, ...)

Examples of supervised approaches

- Regression (Logistic, linear, Cox ...)
- Penalized regression (LASSO, elastic net ...)
- Tree-based methods (CART)
- Ensembles approaches (random forests, gradient boosting ...)
- Neural networks

Examples of unsupervised approaches

- Cluster analysis
- Principal component analysis
- Latent class analysis

3 Penalized Regression

3 Penalized Regression

- Hybrid approaches (aka semi-ML) include both components of machine learning (e.g bootstrap, CV, hyperparameters search) and of statistics (e.g. regression modeling)
- When feasible, these represent ideal setting to get the best out of 2 worlds
- We will discuss penalized regression. Other examples include methods for correlated exposures such as weighted quantile sum (WQS) regression, or other regression approaches such as generalized addiitive models (GAM) that robustly incorporate splines in regression using CV[12]

[12] If interested in additional details, see here

3.1 Introduction to penalized regression

In linear regression models we aim at predicting n observations of the response variable, Y, with a linear combination of m predictor variables, X, and a normally distributed error term with variance σ^2:

$$Y = X\beta + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

We need to estimate the parameters, β, from the sample

Ordinary least square (OLS) approach estimates $\hat{\beta}$ in such a way that the sum of squares of residuals is as small as possible. In other words, we minimize the following loss function:

$$L_{OLS}(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i'\hat{\beta})^2 = \|y - X\hat{\beta}\|^2$$

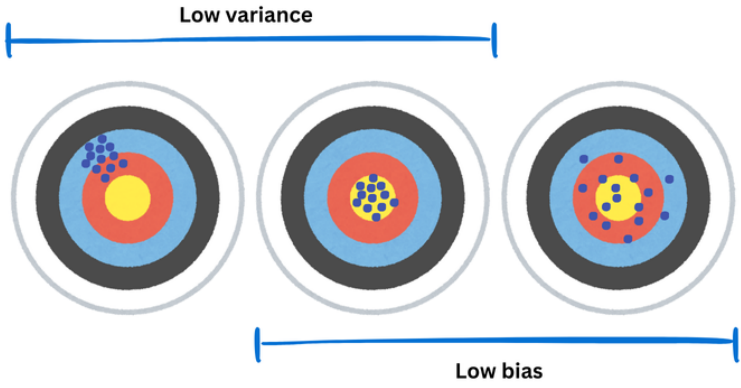
In statistics, there are two critical characteristics of estimators to be considered: the bias and the variance

- The bias measures the accuracy of the estimates:

$$Bias(\hat{\beta}_{OLS}) = E(\hat{\beta}_{OLS}) - \beta$$

- The variance measures the uncertainty of the estimates:

$$Var(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$$

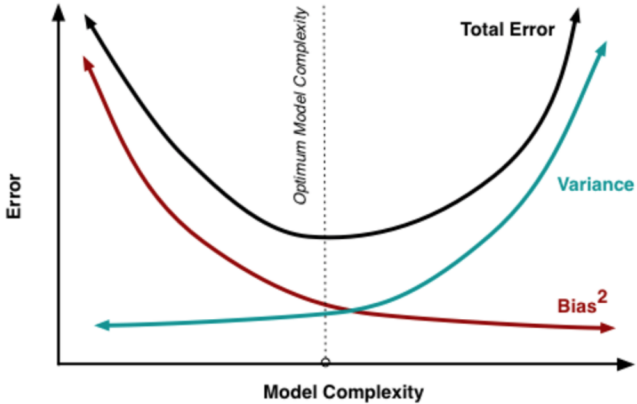

Low variance
Low bias

Variance-bias trade-off

- One of the features that will increase variance is model complexity. Overly complex models (e.g. too may predictors, variable transformations and interaction effects) become more flexible but will tend to better fit and reduce bias in the training data
- In practical terms, by only focusing on bias, we optimize the model for the training set and increase the risk of overfitting
- Under complex settings we therefore need to make sure that minimizing bias does not overly inflate variance (and thus overfitting). This dilemma is known as bias-variance tradeoff in ML literature
- As a common solution, ML algorithms are generally trained to minimize the (Root) Mean Squared Error, which is equal to the sum of Variance and squared Bias

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

Penalized regression: let's reduce the variance at the cost of introducing some bias



Bias and variance as a function of model compelxity

- How do we do that? Through regularization methods where a penalty term is added.
- Penalties are added to the overall estimator and will result in shrinkage of coefficients that are far from 0
- Three common penalty terms and resulting estimating approaches: ridge, LASSO, elastic net

The Ridge penalty, also known as L2 regularization, penalizes OLS based on the square of the magnitude of the coefficients[13]

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} \hat{\beta}_j^2 = \|y - X\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|^2$$

[13] No need to understand the specific of these equations but just to compare them to the OLS estimator presented before

The LASSO (Least Absolute Shrinkage and Selection Operator), also known as L1 regularization, penalizes OLS based on the absolute value of the magnitude of the coefficients

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

- Differently from ridge, coefficients can be shirnked down to 0, thus effectively operating variable selection
- For both ridge and LASSO, the new parameter λ is the regularization or penalty term, When $\lambda = 0$ there is noe penalty and the methods coincide with OLS

Elastic-net combines L1 and L2 regularization

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{m} |\hat{\beta}_j| \right)$$

- The additional parameter α indicates whether we are leaning closer to ridge ($\alpha = 0$) or LASSO ($\alpha = 1$)

How to choose Lambda?

Key step in penalized regression

- As λ becomes larger, the variance decreases and the bias increases
- We use CV to choose the data-driven optimal value that balances bias control and variance (overfitting) control
- Specifically, we are looking for the value of λ that provides the lowest RMSE (λ_{min})
- To improve generalizability, in practice, we tend to focus on the largest value of λ such that error is within 1 standard error of the cross-validated errors for λ_{min}
- Next, after having used CV for identifying the hyperparameter λ, we plug this final value into the penalized regression model

Practical example [more in lab]:

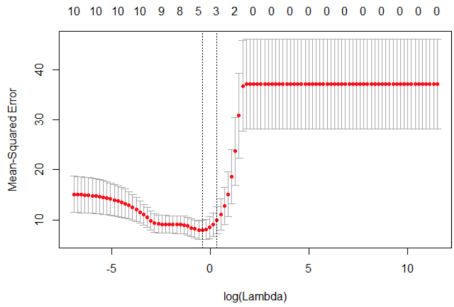
First part (ML): hyperparameter tuning

```
# 1. Define grid for hyperparameters search
lambdas_to_try <- 10^seq(-3, 5, length.out = 100)

# 2. Run CV models for all hyperparameters
lasso_cv <- cv.glmnet(x, y, alpha = 1, lambda = lambdas_to_try,
                      nfolds = 100,family="binomial")

# 3. Identify optimal lambda for balancing precision and overfitting control
lasso_cv$lambda.1se
```

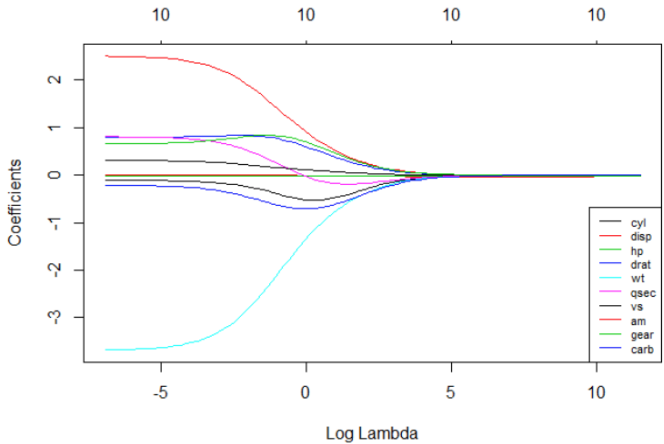
Visual assessment of CV results: RMSE as a function of λ



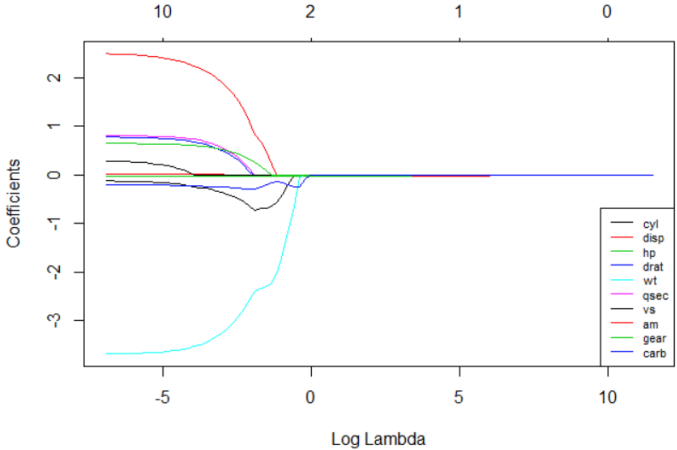
RMSE as a function of Lambda

- On top of the figure we see how many variables are retained in the model when variable selection is operated (LASSO or elastic net)
- The two red lines indicate the min and the cross-validated λ

It is also possible to visualize how the individual coefficients change as a function of the penalization. Here an example with ridge, where coefficients are shrunk towards 0 but never removed from the model:



And an example with LASSO, where coefficients can be shrunk to 0 (variable selection)



Second part (regression): fit the final model using the selected penalty

```
model_cv_lasso <- glmnet(x, y, alpha = 1, lambda = lasso_cv$lambda.1se)
```

- Visual output (examples to follow) include list of selected predictors or estimates of regression coefficients if relevant

LASSO/Elastic Net: additional notes

- LASSO is a penalized regression approach that performs variable selection. For this reason, it is more commonly used in clinical/observational epidemiology as compared to Ridge
- Elastic Net can offer advantages over LASSO in specific settings:
  - When operating variable selection, it offers a less conservative approach (i.e. more variables are included)
  - With highly correlated predictors, EN is more efficient than LASSO in identifying the correct driver of associations
- When using EN, the hyperparameter search can be conducted for joint levels of α and λ
- In epi applications, it is common to proceed in 2 stages:
  - LASSO/EN for variable selection robust to complex settings such as high correlations
  - Incorporate selected variables in un-penalized regressions

Penalized regression in clinical epidemiology

- LASSO offers a powerful tool for variable selection in clinical prediction modeling, providing considerable advantages over standard procedures such as stepwise regression[14]
- Some of its advantages include:
  - Robustly addressing overfitting during the selection procedure (CV) rather than after (training/validation split)
  - Direct selection focus on AIC and RMSE rather than Wald test or p-values
  - It can work with p>n (yet, beware of sparse data)
  - It deals better with correlated predictors
- In practice, one could standard with backward selection and validate results with LASSO

[14] See here and here

Example 1

JAMA Internal Medicine

Home | Issues | Multimedia

Home | JAMA Internal Medicine | Vol. 180, No. 8

Original Investigation FREE

Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19

Wenhua Liang, MD[1,2]; Hengrui Liang, MD[1,2]; Limin Ou, MD[1]; et al

» Author Affiliations | Article Information

Variable Selection and Score Construction

All 1590 patients hospitalized with COVID-19 in the development cohort were included for variable selection and risk score development. As described herein, 72 variables were entered into the selection process. Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied to minimize the potential collinearity of variables measured from the same patient and over-fitting of variables. Imputation for missing variables was considered if missing values were less than 20%. We used predictive mean matching to impute numeric features, logistic regression to impute binary variables, and Bayesian polytomous regression to impute factor features. We used L1-penalized least absolute shrinkage and selection regression for multivariable analyses, augmented with 10-fold cross validation for internal validation. This is a logistic regression model that penalizes the absolute size of the coefficients of a regression model based on the value of λ. With larger penalties, the estimates of weaker factors shrink toward zero, so that only the strongest predictors remain in the model. The most predictive covariates were selected by the minimum (λ min). The R package "glmnet" statistical software (R Foundation) was used to perform the LASSO regression. Subsequently, variables identified by LASSO regression analysis were entered into logistic regression models and those that were consistently statistically significant were used to construct the risk score (COVID-GRAM),[7] which was then used to construct a web-based risk calculator (http://118.126.104.170/). Data were analyzed between February 20, 2020 and March 17, 2020.

Table 3. Multivariable Logistic Regression Model for Predicting Development of Critical Illness in 1590 Patients Hospitalized With COVID-19 in Wuhan

Variables	Odds ratio (95% CI)	*P* value
X-ray abnormality (yes vs no)	3.39 (2.14-5.38)	<.001
Age, per y	1.03 (1.01-1.05)	.002
Hemoptysis (yes vs no)	4.53 (1.36-15.15)	.01
Dyspnea (yes vs no)	1.88 (1.18-3.01)	.01
Unconsciousness (yes vs no)	4.71 (1.39-15.98)	.01
No. of comorbidities	1.60 (1.27-2.00)	<.001
Cancer history (yes vs no)	4.07 (1.23-13.43)	.02
Neutrophil to lymphocyte ratio	1.06 (1.02-1.10)	.003
Lactate dehydrogenase, U/L	1.002 (1.001-1.004)	<.001
Direct bilirubin, μmol/L	1.15 (1.06-1.24)	.001
Constant	0.001	

Abbreviation: COVID-19, coronavirus disease 2019.

Example 2

▸ BMJ Ment Health. 2023 Aug 21;26(1):e300719. doi: 10.1136/bmjment-2023-300719

Development and validation of a dementia risk score in the UK Biobank and Whitehall II cohorts

Melis Anatürk [1,2,0], Raihaan Patel [2,3,✉,0], Klaus P Ebmeier [2], Georgios Georgiopoulos [4], Danielle Newby [2], Anya Topiwala [2,5], Ann-Marie G de Lange [2,6,7], James H Cole [1,8], Michelle G Jansen [9], Archana Singh-Manoux [10,11], Mika Kivimäki [11], Sana Suri [2,3]

Statistical analyses: development of the UKBDRS

All continuous variables were standardised and outliers (ie, individuals with values <Q1 − 3×IQR or values >Q3 + 3×IQR) were excluded (1.8% of dataset). The first stage ('variable selection') used only the training set and involved submitting the 28 candidate predictors to a least absolute shrinkage and selection operator (LASSO) Cox regression (online supplemental methods), to identify a parsimonious model with dementia as the outcome.[18] Correlation between numerical variables was assessed before LASSO (online supplemental methods, online supplemental figure 2). LASSO selected variables were then used as predictors in a Fine and Gray competing risk regression model.[19] Duration of follow-up was calculated as time between baseline and either date of dementia, death, or censoring date (online supplemental methods). The linear predictor was used to compute the predicted probability of developing dementia (online supplemental methods). Two variants of the

Note: even when using LASSO, it is still better to develop the CV-model within a training set. Overfitting is addressed by both CV and training/validation split

Selection of predictors for the UKBDRS

LASSO regression identified 11 variables as predictive of incident dementia: age, education, history of diabetes, history/current depression, history of stroke, parental history of dementia, Townsend deprivation, hypertension, high cholesterol, household occupancy (living alone), and sex (online supplemental table 5). The beta coefficients for the final competing risk regression models are provided in table 1.

Table 1. Results of the competing risk regressions with two variants of the UKBDRS.

Predictor	β	HR	95% CI Lower	Upper	P
UKBDRS					
Age (years)	0.178	1.194	1.184	1.206	2.1×10^{-296}
Parental history (yes)	0.431	1.539	1.415	1.674	2.1×10^{-296}
Education (years)	−0.041	0.960	0.948	0.972	2.1×10^{-296}
Townsend deprivation (most deprived)	0.228	1.256	1.153	1.367	2.1×10^{-296}
Diabetes (yes)	0.536	1.710	1.528	1.914	2.1×10^{-296}
Depression (yes)	0.556	1.744	1.593	1.909	2.1×10^{-296}
Stroke (yes)	0.655	1.925	1.652	2.242	2.1×10^{-296}
Hypertensive (yes)	0.159	1.173	1.082	1.271	2.1×10^{-296}
High cholesterol (yes)	0.104	1.110	1.015	1.213	2.1×10^{-296}
Sex (male)	0.169	1.184	1.099	1.275	2.3×10^{-2}
Lives alone (yes)	0.141	1.151	1.058	1.253	1×10^{-3}

Penalized regression in etiological epidemiology

- LASSO allows for variable selection when complex and possibly high-predictors predictors are present
- One of its advantages is that it robustly deals with correlated predictors by (usually) selecting variables truly associated with the outcome out of a correlated cluster
- Differently from clinical prediction, where parsimony is key, one might not want to be too conservative in the selection. Hence, elastic net is often used as primary modeling choice
- These advantages are particularly relevant when working with complex exposures of environmental mixtures (e.g PM2.5, metals, or even nutrients)

Example 1

A Section 508–conformant HTML version of this article
is available at http://dx.doi.org/10.1289/ehp.1408933.

Prenatal Phthalate, Perfluoroalkyl Acid, and Organochlorine Exposures and Term Birth Weight in Three Birth Cohorts: Multi-Pollutant Models Based on Elastic Net Regression

Virissa Lenters,[1] Lützen Portengen,[1] Anna Rignell-Hydbom,[2] Bo A.G. Jönsson,[2] Christian H. Lindh,[2] Aldert H. Piersma,[3] Gunnar Toft,[4] Jens Peter Bonde,[5] Dick Heederik,[1] Lars Rylander,[2] and Roel Vermeulen[1,6*]*

[1]Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands; [2]Division of Occupational and Environmental Medicine, Lund University, Lund, Sweden; [3]Laboratory for Health Protection Research, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands; [4]Danish Ramazzini Center, Department of Occupational Medicine, Aarhus University Hospital, Aarhus, Denmark; [5]Department of Occupational and Environmental Medicine, Copenhagen University Hospital, Bispebjerg, Copenhagen, Denmark; [6]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

- Co-exposure to 16 chemicals
- Birth weight in 1250 infants

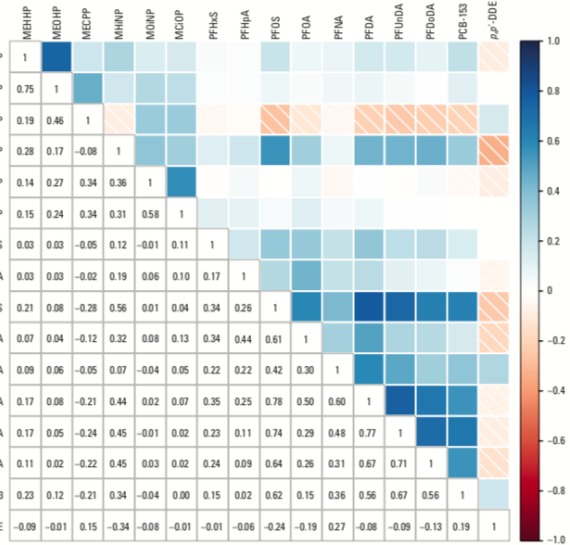


Figure 2. Spearman correlation coefficients between exposure biomarkers. The color intensity of shaded boxes indicates the magnitude of the correlation. Blue indicates a positive correlation, and red with white diagonal lines indicates a negative correlation.

From Statistical Analysis Section:

"*In the case of multicollinearity, multiple linear regression models may yield unreliable parameter estimates. Therefore, to assess which exposures are associated with the outcome while simultaneously adjusting for other exposures, we used elastic net regression modeling [. . .] Whereas ridge retains all predictors, and lasso tends to select only one predictor from a group of correlated predictors, elastic net can perform selection while enabling the inclusion of collinear predictors in the final model.*''

"*To fully adjust for potential confounders, we selected confounders a priori, without subjecting them to variable selection, and included them in the elastic net regression models as unpenalized variables.*''

"*We used cross-validation (CV) to determine the optimal degree of penalization. We tested models over a grid of α and λ sequences, and selected the combination which yielded the minimum mean-squared error (MSE) of prediction from repeated 10-fold CV. CV was repeated 100 times, each time with different data partitions, to achieve more stable selection than a single CV.*''

"*For the subset of exposures selected via elastic net [...] we refit multiple-exposure ordinary least squares (OLS) regression models to obtain unpenalized, mutually adjusted coefficient estimates.*''

"*Before modeling, exposure variables were natural log-transformed to reduce the influence of outliers. Data below the LOD (0-27%) were singly imputed [...] For regression models, we mean-centered all predictor variables and scaled continuous variables by two times their respective standard deviations (SD)*''

Results

Table 2. Multiple-exposure elastic net penalized regression models[a] (β_{EN}) for term birth weight.

Potential predictor (increment)	Adjusted	Plus gestational age	Further adjusted
ln-MEHHP (1.70 ng/mL)	–64.67[b]	–59.43[b]	–48.61
ln-MEOHP (1.29 ng/mL)	–0.15	0	0
ln-MECPP (1.42 ng/mL)	0	0	0
ln-MHiNP (2.74 ng/mL)	0	0	0
ln-MOiNP (2.22 ng/mL)	23.81	22.26	16.31
ln-MCiOP (2.32 ng/mL)	0	0	0
ln-PFHxS (1.24 ng/mL)	–3.49	0	0
ln-PFHpA (1.84 ng/mL)	0	0	0
ln-PFOS (1.60 ng/mL)	0	0	0
ln-PFOA (1.18 ng/mL)	–11.51	–10.11	–38.82
ln-PFNA (1.03 ng/mL)	–7.05	–7.69	0
ln-PFDA (1.40 ng/mL)	0	0	0
ln-PFUnDA (2.10 ng/mL)	0	0	0
ln-PFDoDA (1.67 ng/mL)	–22.56	0	0
ln-PCB-153 (2.43 ng/g)	0	0	0
ln-*p,p*′-DDE (1.82 ng/g)	–106.39[b]	–76.63[b]	–47.02[b]

Regression coefficients (β_{EN}) represent the change in mean birth weight (g) for term infants per increment: a 2-SD increase in ln-transformed exposure biomarker levels. β_{EN} for the modeled, unpenalized covariates are not shown.
[a]The cross-validated optimum penalization was $\alpha = 1.00$, $\lambda = 3.32$ (MSE = 205,061) for the adjusted model (minimal sufficient set: study population, maternal age, prepregnancy BMI, and parity); $\alpha = 1.00$, $\lambda = 3.32$ (MSE = 177,179) for the model additionally adjusted for gestational age; and $\alpha = 0.98$, $\lambda = 2.46$ (MSE = 166,329) for the further adjusted model (plus infant sex, maternal height, alcohol consumption, serum cotinine, and vitamin D). All models, $n = 1{,}250$. [b]Covariance test

Table 3. Multiple-exposure unpenalized linear regression models for the exposures selected via elastic net regression and term birth weight [β_{OLS} (95% CI)].

Predictor (increment)	Adjusted	*p*-Value	Plus gestational age	*p*-Value	Further adjusted	*p*-Value
ln-MEHHP (1.70 ng/mL)	−86.75 (−139.18, −34.32)	0.001	−83.94 (−132.68, −35.19)	0.001	−70.22 (−117.59, −22.85)	0.004
ln-MOiNP (2.22 ng/mL)	45.85 (−4.84, 96.54)	0.076	45.62 (−1.51, 92.74)	0.058	37.64 (−7.99, 83.27)	0.106
ln-PFOA (1.18 ng/mL)	−42.77 (−108.19, 22.65)	0.200	−41.02 (−101.83, 19.80)	0.186	−63.77 (−122.83, −4.71)	0.035
ln-*p,p*′-DDE[a] (1.82 ng/g)	−134.73 (−191.93, −77.53)	< 0.001	−100.75 (−154.13, −47.36)	< 0.001	−66.70 (−119.38, −14.02)	0.013
Population						
Poland	−40.24 (−133.58, 53.11)	0.398	4.84 (−82.16, 91.85)	0.913	−93.16 (−183.72, −2.60)	0.044
Ukraine	−218.89 (−300.66, −137.13)	< 0.001	−142.98 (−219.73, −66.22)	< 0.001	−256.90 (−338.77, −175.02)	< 0.001
Maternal age (years)						
27–31	88.14 (23.05, 153.24)	0.008	80.01 (19.48, 140.53)	0.010	65.44 (6.74, 124.14)	0.029
32–45	30.53 (−42.52, 103.57)	0.413	38.08 (−29.83, 105.99)	0.272	37.19 (−28.83, 103.21)	0.270
BMI (8.62 kg/m^2)	209.00 (155.53, 262.46)	< 0.001	181.37 (131.51, 231.22)	< 0.001	194.00 (145.42, 242.58)	< 0.001
Parity: multiparous	58.72 (−3.63, 121.06)	0.065	76.33 (18.32, 134.34)	0.010	85.92 (29.39, 142.44)	0.003
Gestational age (2.45 weeks)			343.78 (295.61, 391.94)	< 0.001	330.43 (283.70, 377.15)	< 0.001
Infant sex: female					−115.40 (−160.38, −70.43)	< 0.001
Maternal height (12.93 cm)					135.83 (88.38, 183.28)	< 0.001
Alcohol: ≥ 7 drinks/week					34.43 (−61.76, 130.62)	0.483
Cotinine (113.51 ng/mL)					−140.41 (−191.92, −88.89)	< 0.001
Vitamin D (22.05 ng/mL)					18.77 (−29.39, 66.93)	0.445

Regression coefficients (β_{OLS}) represent the change in mean birth weight (g) for term infants per increment: a 2-SD increase in ln-transformed exposure biomarker or untransformed continuous covariate levels, or per category for categorical covariates. Reference categories are population, Greenland; maternal age, 18–26 years; parity, nulliparous; infant sex, male; alcohol, < 7 drinks/week (around the time of conception). Variance inflation factors for exposure terms ranged from 1.04 to 1.74.
[a]β_{OLS} for models including wet weight *p,p*′-DDE (ng/mL), adjusted for total lipids: −134.22 (95% CI: −191.43, −77.02), −99.91 (95% CI: −153.30, −46.52), −67.16 (95% CI: −119.80, −14.51).

Example 2

Environment International 106 (2017) 60–68

Contents lists available at ScienceDirect

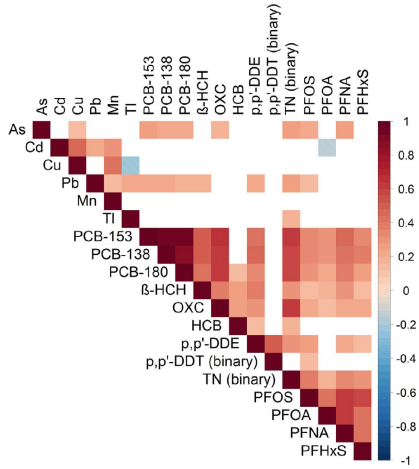
Environment International

journal homepage: www.elsevier.com/locate/envint

Neonatal exposure to environmental pollutants and placental mitochondrial DNA content: A multi-pollutant approach

Annette Vriens[a], Tim S. Nawrot[a,b], Willy Baeyens[c], Elly Den Hond[d], Liesbeth Bruckers[e], Adrian Covaci[f], Kim Croes[c], Sam De Craemer[c], Eva Govarts[g], Nathalie Lambrechts[g], Ilse Loots[h], Vera Nelen[d], Martien Peusens[a], Stefaan De Henauw[i,j], Greet Schoeters[g], Michelle Plusquin[a,*]

- Co-exposure to 4 perfluoroalkyl compounds and 9 organochlorine compounds
- mtDNA content in placental tissue of 233 newborns


As
Cd
Cu
Pb
Mn
Tl
PCB-153
PCB-138
PCB-180
ß-HCH
OXC
HCB
p,p'-DDE
p,p'-DDT (binary)
TN (binary)
PFOS
PFOA
PFNA
PFHxS
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

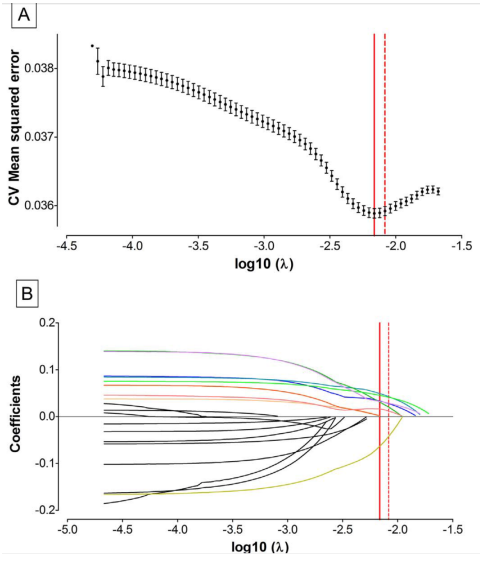
From Statistical Analysis Section

"*The relevant biomarkers of environmental exposures were selected using least absolute shrinkage and selection operator (LASSO) penalized regression analysis. This method allows selecting the important predictors and also estimates the regression coefficient. However, it should be noted that the effect estimates of LASSO are biased as a consequence of the penalty*''

"*The optimal value of the penalty parameter λ was determined using 10-fold cross-validation. The LASSO model was tested over a grid of λ values and the λ where the mean squared error (MSE) was within one standard error of the minimum MSE of prediction from 100 times repeated cross-validation was chosen.*''

"*The independent variables (i.e. the exposures and covariates) were standardized prior to fitting the LASSO regression model. We applied multiple ordinary least squares (OLS) regression models to estimate the adjusted effects of the exposures, which were selected by LASSO.*''

Results


A
CV Mean squared error
0.038
0.037
0.036
-4.5
-4.0
-3.5
-3.0
-2.5
-2.0
-1.5
log10 (λ)
B
Coefficients
0.2
0.1
0.0
-0.1
-0.2
-5.0
-4.5
-4.0
-3.5
-3.0
-2.5
-2.0
-1.5
log10 (λ)

"*Cross-validation selected the λ with the lowest MSE and the highest λ within one SE of the minimum MSE (Fig. 2A, respectively indicated by the full red line and dashed red line). For highest λ within 1 SE of the minimum MSE (dashed red line in Fig. 2B), OXC, β-HCH, p,p'-DDE, PFNA, As, Cd and Tl, were selected as important predictors for placental mtDNA content, with non-zero coefficients.*''

Table 3
The estimated effects of environmental pollutants on placental mtDNA content.

Pollutant	Estimated effect (95% CI)	*p*-value
Oxychlordane	1.03 (− 1.29, 3.41)	0.39
β-hexachlorocyclohexane	2.71 (0.25, 5.23)	0.0307
p,p′-dichlorodiphenyldichloroethylene	0.99 (− 0.80, 2.80)	0.28
Arsenic	1.41 (0.07, 2.77)	0.0389
Cadmium	2.51 (− 0.64, 5.75)	0.12
Thallium	− 4.88 (− 9.09, − 0.48)	0.0303
Perfluorononanoic acid	1.07 (− 0.94, 3.12)	0.30

The estimated effects of the pollutants were determined using a multiple OLS regression model. The estimated effects are represented as a percent (%) change in placental mtDNA content for a 25% increment in mean concentration of the environmental pollutants in cord blood. The model was adjusted for sex, gestational age, season of birth, maternal age, maternal pre-pregnancy BMI, smoking during pregnancy, parity and education. Gestational age, maternal pre-pregnancy BMI, smoking during pregnancy and season of birth were significantly associated with placental mtDNA content in our model.

4.Supervised Machine Learning

4.Supervised Machine Learning

- With full-scale machine learning, we remove all regression-based assumption and let the machine work out all sections of the analyses without explicit assumptions[15]
- We will mostly focus on ensemble methods
- These build up on some variations of tree-based approaches such as classification and regression trees

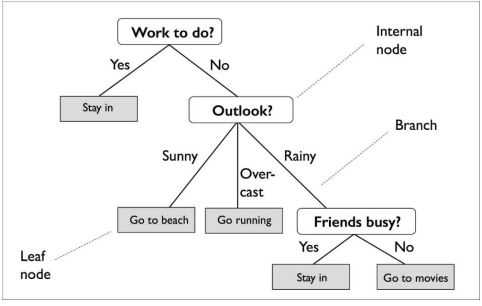
[15] Note that ML is not completely assumptions-free. For example, most methods require data to be independent and identically distribute, assumptions that might fail in the presence of serial data or clustered individuals

4.1 Classification and Regression Trees (CART)

- Classical ML approaches introduced by Breiman in the 1980s
- Based on recursive splitting the data
- Target variable
  - Probability (0/1) -> classification
  - Continuous endpoint -> regression
- Based on the intuitive idea of recursive partitioning

Recursive partitioning

Example[16]

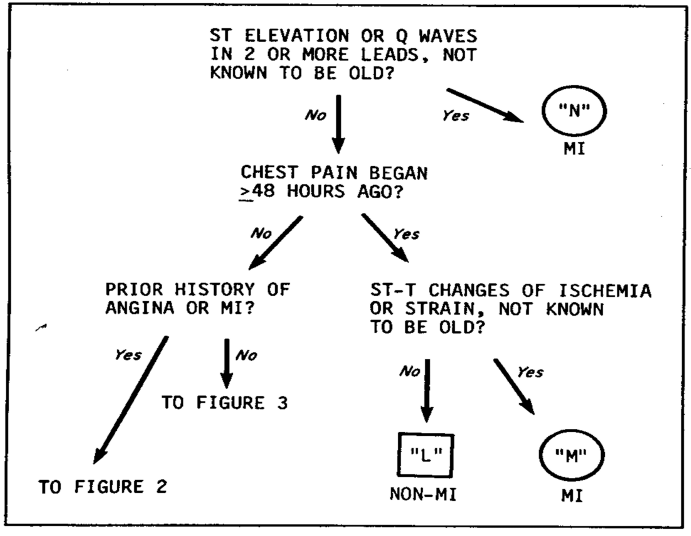


- Identify the first split
- if needed, split again within groups

[16] From here

Pre-ML example of recursive partitioning in clinical research with

- N Engl J Med 1988;318:797-803.
- 1379 patients who presented at an emergency room with the chief complaint of chest pain
- Outcome: Myocardial Infarction
- Predictors: History, Signs/Symptoms, Test Results

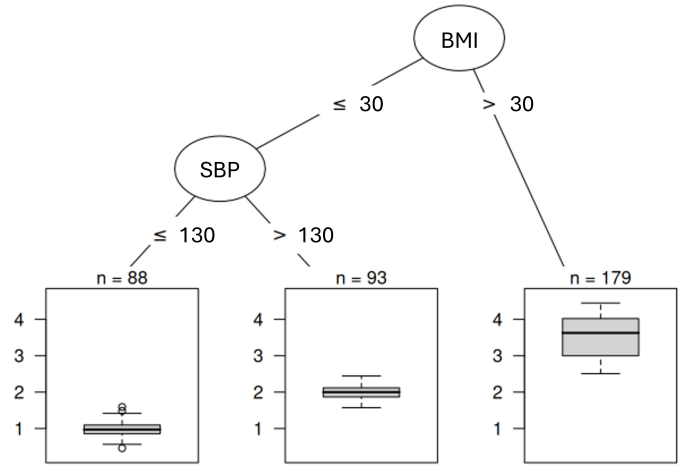


From NEJM 1988

Tree-based models

- Say we are interested in predicting the risk of MI based on several patients characteristics. We incorporate continuous covariates (e.g. BMI, SBP) and we want a selection procedure that does not make any assumption of linearity or additivity for the relationships between the potential predictors and MI
- Tree-based models split the data multiple times according to specific cutoffs of the different variables (called *features* in ML terminology)
- These splits are generated with the goal of creating subgroups as different as possible with respect to the target outcome (aka *label*)

Example:



How are non-linear and non-additive effects integrated?

- A variable can appear more than once in the splits
- This, in a nutshell, is how tree-based methods handles non-additive and non-linear effects
- For example, split BMI>30 vs BMI<30 as a feature for high vs low risk in a first step
- In a following step, split BMI <25 vs BMI >30 as a feature for high vs low risk
- The machine is telling us that BMI<25 and BMI>30 are both higher risk than 25<BMI<30 (non-lienar effect)
- Moreover, the recursive nature implicitely allows to incorporate effect modification (i.e. non-additivity). In the example before, for example, the effect of BMI<30 varies depending on the value of SBP

Classification and regression trees

- CART are one type of tree-based methods[17]
- Specifically, CART takes a feature and determines which cut-off point minimizes the variance of the outcome for a regression task or the Gini index[18] for classification tasks.
- The algorithm continues this search-and-split recursively in both new nodes until a stop criterion is reached:
  - A minimum number of individuals have to be in a node before the split
  - A minimum number of individuals have to be in a terminal node.

[17] See The Elements of Statistical Learning for more details

[18] A measure of how much unequal the values of a frequency distribution are between subgroups

Other supervised methods

Several other methods exist and have been presented also in epi contexts. Two methods worth mentioning, not covered here, include:[19]

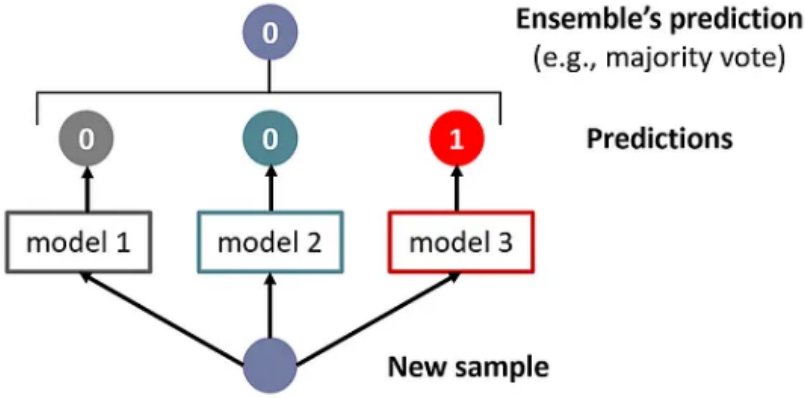
- Support Vector Machines
- Naive Bayes Algorithms

[19] See here for example

4.2 Ensemble methods

- CART do not offer robust control for overfitting even when accurate data splits are used
- The final prediction (e.g. MI yes vs no, going out yes vs no) can be subject to the specific chosen model

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model


0
Ensemble's prediction
(e.g., majority vote)
0
0
1
Predictions
model 1
model 2
model 3
New sample

Main approaches are:

- Bagging
- Random forests
- Gradient boosting

We'll briefly discuss bagging and random forests, and focus more extensively on boosting

Bagging

Bootstrap Aggregating

- Take repeated bootstrap samples of data
- Generate prediction rules on each sample
- Final prediction = average of prediction rules from each sample

Bagging algorithm:

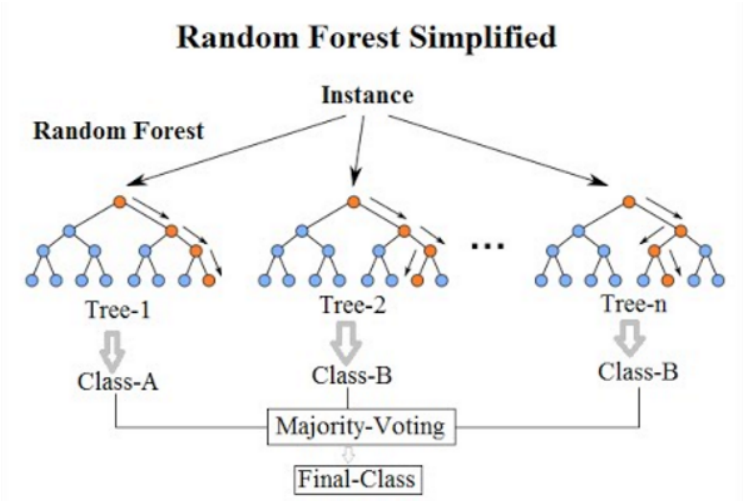
- Draw K bootstrap Samples $B_1, \ldots, B_K$ with replacement from data set
- Each bootstrap sample is of same size as data set
- Construct K trees $T_1, \ldots, T_K$ from the bootstrap samples
- For each person, estimate probability or class from each tree
- Average probability estimates over trees (or use majority vote)

Random forests

- The fundamental idea behind a random forest is to combine many decision trees into a single model
- Similar to having hundreds of humans making a guess and trusting the average better than the prediction of a single individual
- Key distinction from bagging is that each decision tree in the forest only has access to a subset of features (predictors)
- This is done to remove noise and redundant information, thus reducing overfitting. Moreover, it increases speed and efficiency

Random forests algorithm

- Draw K bootstrap Samples $B_1, \ldots, B_K$ with replacement from data set
- Construct K trees $T_1, \ldots, T_K$ from the bootstrap samples
  - Randomly select m variables for each potential split and tree
  - Select best split from these m variables
  - K is typically high, well above 100
- Average probability estimates over trees (or use majority vote)


Random Forest Simplified
Instance
Random Forest
Tree-1
Tree-2
Tree-n
Class-A
Class-B
Class-B
Majority-Voting
Final-Class

Tuning parameters in a random forest

- All those steps are the machine's task
- Humans tasks are:
  - To identify the candidate predictors or, in general, the variables you'll feed to the machine[20]
  - To define the hyperparameters for the tuning phase
- In a random forest, key hyperparameters include:
  - Size of Forest (how many trees)
  - Number of candidate splits at each node of each tree (often SQRT(# predictors), or $2SQRT$ / $.5$SQRT)
  - The maximum depth of each tree

[20] we'll get back to this point, but note the difference from a naive approach of feeding all possible variables

Boosting

Hastie et al:[21] "Boosting is one of the most powerful learning ideas introduced in the last 10 years"

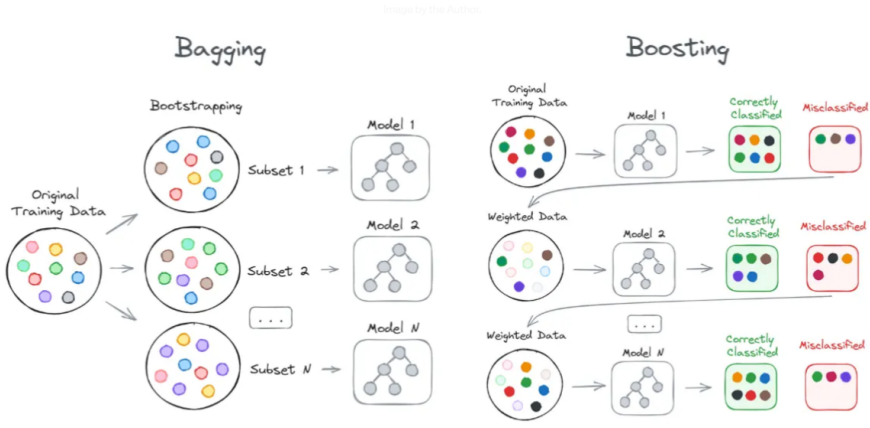
Similar concept to random forest (ensemble of trees) but with a different learning procedure

[21] Authors of the seminal textbook Elements of Statistical Learning

- Training is sequential. Trees are modeled sequentially, with each model trying to correct errors made by the previous ones
- Models can be weighted based on their performances
- The process of adaptive learning minimizes errors and improves accuracy

1. Train a tree to predict probability of the outcome P(Outcome)
2. Instead of training a second independent tree, calculate the residuals of the first tree (i.e. how off you were in your prediction)
3. Build a regression tree to predict the residuals
4. Update the prediction based on this new output
5. Calculate the new residual
6. Repeat usually 3-5 times

Bagging (and random forests) vs boosting[22]



[22] from this article

AdaBoost (Adaptive Boosting)

- One of the earliest boosting algorithms
- Trains a series of simple models (*weak learners*, like decision trees with a single split) and improves over these
- Misclassified data points are provided higher weights at subsequent iterations
- Final prediction provided by weighted sum of individual model's predictions

Gradient Boosting Machines (GBM)

- Similar process to AdaBoost
- Goal is to optimize a well-defined loss function. **This property makes it of particular appeal for epi applications as the same loss functions of classical methods (e.g. linear, logistic, Cox) can be integrated**
- It uses a gradient descent estimation procedure[23]
- The goal function used by subsequent models is the residuals of the previous models

[23] Details beyond the scope of the course

Extreme Gradient Boosting (XGBoost)

- Optimized and scalable performance of GBM
- Advantages in high-dimensional settings where speed and versatility are required
- This added versatility includes:
  - Incorporating L1 and L2 regularization
  - Handling missing values during training

Additional boosting algorithms:

- Stochastic Gradient Boosting (SGB)
- LightGBM
- CatBoost (Categorical Boosting)

Boosting Hyperparameters

Core hyperparameters in boosting training include:

1. Learning (or shrinkage) rate (aka *eta*)

- Between 0 and 1, it controls the size of the step during each iteration
- After each tree is trained, its predictions are shrunk by multiplying them by *eta*
- Lower levels of shrinkage will result in stricter overfitting control

2. Depth

- The maximum depth of each tree
- This is a very important parameter as it connects to the idea of interaction
- For example, depth=2 corresponds to accounting for 2-ways interactions[24]

3. Number of iterations and n of trees
4. Size of each tree

[24] Think of the simple case of 2 binary covariates: a tree with depth=2 will split one of the 2 variables and then each group based on levels of the second variable. This corresponds to looking across the 4 possible strata without any additivity assumption

Gradient boosting, practical steps[25]

1. Define training and validation splits
2. Identify a grid search for all hyperparameters (i.e. all combinations of parameters you want to check: shrinkage 0.1, 0.5, 0.9..., depths=1,2,3..)
3. CV for hyperparameters tuning

[25] Example from this article. More details in lab session

Example

```
```{r}

##    shrinkage interaction.depth optimal_trees  min_RMSE
## 1       0.01                 5          3867  16647.87
## 2       0.01                 5          4209  16960.78
## 3       0.01                 5          4281  17084.29
## 4       0.10                 3           489  17093.77
## 5       0.01                 3          4777  17121.26
## 6       0.01                 3          4919  17139.59
## 7       0.01                 3          4997  17139.88
## 8       0.01                 5          4123  17162.60
## 9       0.01                 5          4850  17247.72
## 10      0.01                 3          4794  17353.36
```
```

4. Train top model [lowest RMSE] in all training set
5. Predict in validation

4.4. Interpretable Machine Learning

- The algorithms that we have described were all designed with the goal of prediction
- The inner mechanisms of the algorithms (i.e. how the machine came to a certain decision) is unknown
- This property is commonly referred to as black box
- Interpretable (aka explainable) ML attempts to get inside this black box and decribe the inner mechanisms

- This step is extremely important in epi and clinical research as the inner mechanisms are often the main goal of our research
- Even when the goal is prediction (e.g. clinical prediction modeling) we are interested in additional details such as what predictors are driving the predictions, what interactions are at play etc.
- In etiological research, on the other hand the mechanism are generally the main target (e.g. identifying real risk factors out of several correlated environmental exposures)

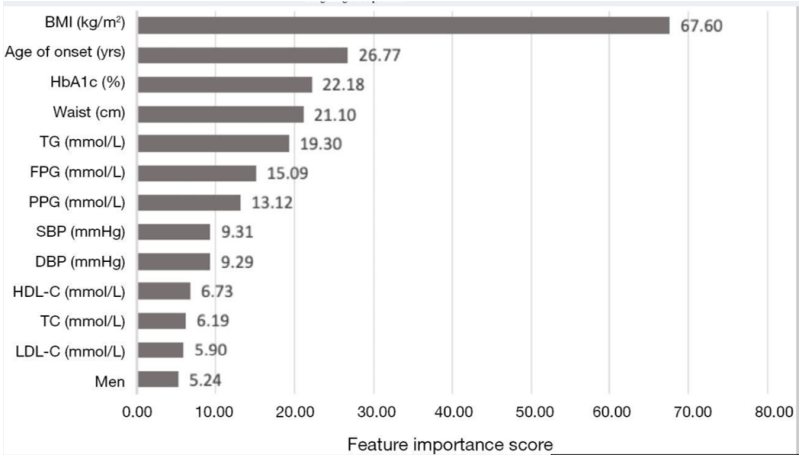
Overview of major interpretable methods

- Interpretable by design
  - Methods such as a single decision rule, logistic regression, or GAM, are inherently interpretable
- Methods for post-hoc interpretability
  - We will focus on a set of model-based model-agnostic methods that can be used to explain results of bagging and boosting
- Additional details in the excellent (and free) book by Molner, available here

Variable (feature) importance

- Most methods provide model-based feture importance. At each split in each tree, they compute the improvement in the split-criterion (eg RMSE) and then average the improvement made by each variable across all the trees where the variable is used
- The variables with the largest average decrease in MSE are considered most important
- Usually presented as a ranking of top contributors
- Also available as a model-agnostic methods (permutation feature importance)
- SHAP (SHapley Additive Explanations) values are another way to quantify the contribution of each feature of individual predictions

Example: results presentation from Tang et al, 2021, Annals of Translational Medicine[26]

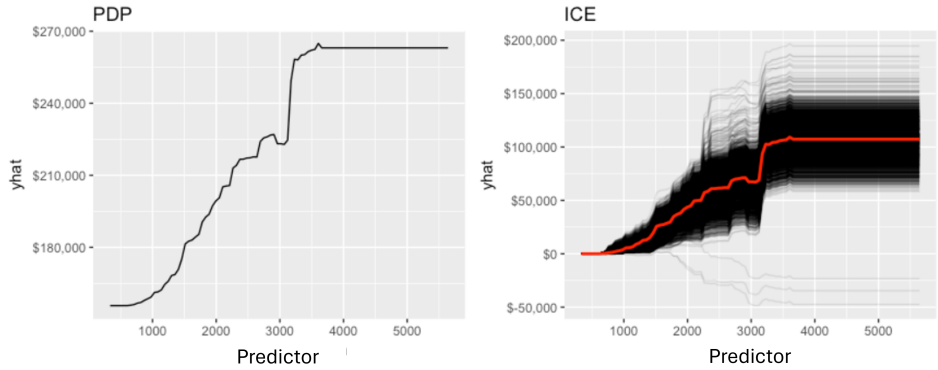


Variables importance

[26] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8033361/

Partial dependence plot

- How is the response variable changing at varyting levels of the relevant predictors
- Useful to identify shape of X-Y associations (e.g. linear vs not)
- Alternative methods include individual conditional expectation (ICE) curves

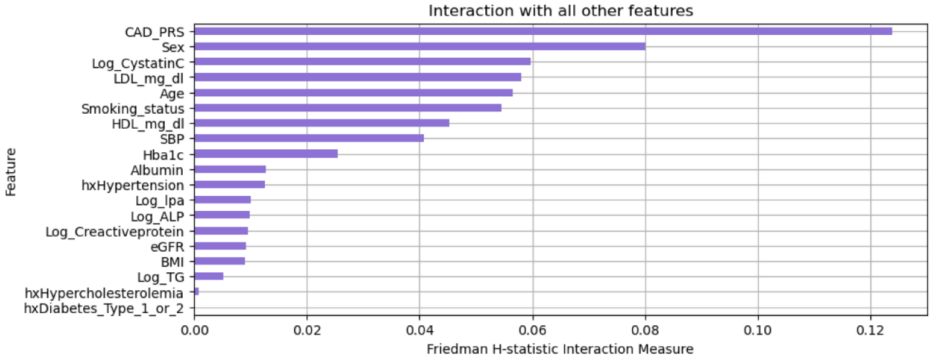


Partial dependence plots

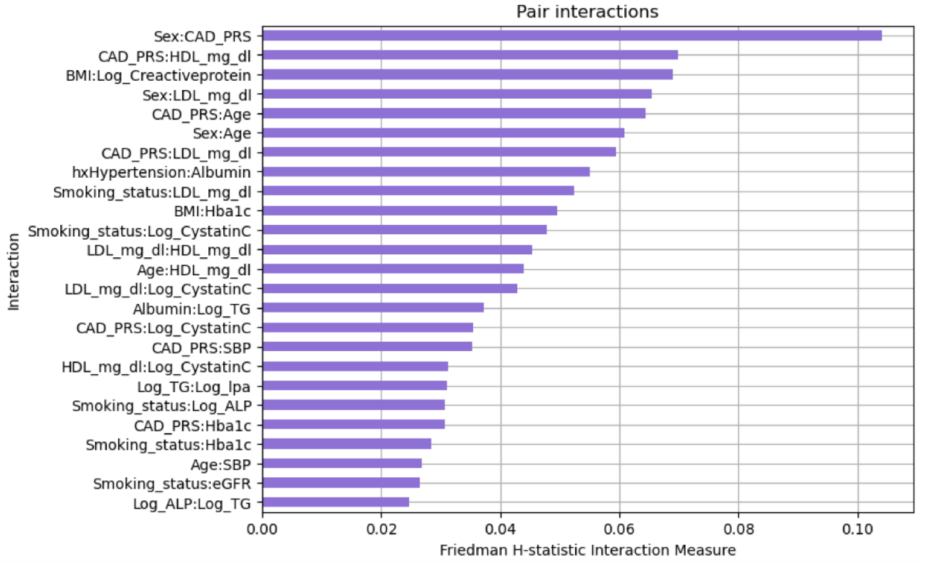
Features interactions

- Calculated using Friedman's H-statistic, which indicate the share of variance that is explained by interactions
- Can be calculated in terms of pairwise interactions or in terms of overall interactions
- Could also be used to analyze higher-order interactions such as the interaction strenght between 3 or more variables
- It is computationally extensive
- Not (yet) available as a model-agnostic version but as a post-hoc estimation after boosting

Example: overall interaction strenght



Example: pairwise interactions



4.5 Applications of supervised ML in epidemiology

Several settings where supervised ML has been explored and applied, including:

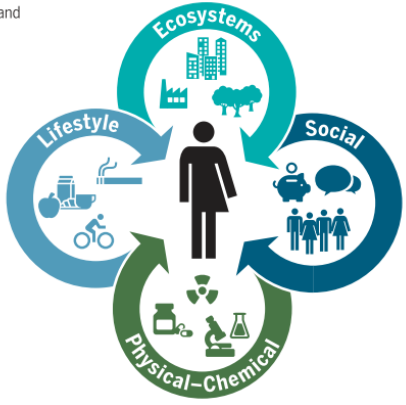
- Exposome-wide analysis
- Refining clinical prediction models

a) Exposome-wide analysis

Ecosystems
Food outlets, alcohol outlets
Built environment and urban land uses
Population density
Walkability
Green/blue space

Lifestyle
Physical activity
Sleep behavior
Diet
Drug use
Smoking
Alcohol use

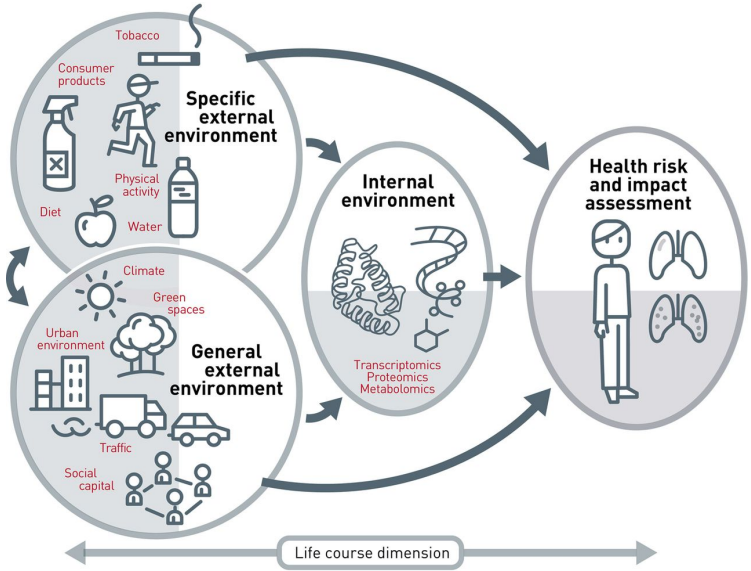
Social
Household income
Inequality
Social capital
Social networks
Cultural norms
Cultural capital
Psychological and mental stress



Physical–Chemical
Temperature/humidity
Electromagnetic fields
Ambient light
Odor and noise
Point, line sources, e.g. factories, ports
Outdoor and indoor air pollution
Agricultural activities, livestock
Pollen/mold/fungus
Pesticides
Fragrance products
Flame retardants (PBDEs)
Persistent organic pollutants
Plastic and plasticizers
Food contaminants
Soil contaminants
Drinking water contamination
Groundwater contamination
Surface water contamination
Occupational exposures

The exposome

The totality of external expoure over the liferime as a determinant , through complex internal mechanisms, of the health profile and risk of an individual



- ML can be used to incorporate several groups of exposures to be explored as potential determinants of health. Several equally complex research questions:
  - Screen for potential associations out of complex datasets
  - Integrate multiple exposures while accounting for co-confounding, non-linearities, and interactions
  - Assess causal effects

Example 1. Rotem et al. 2024


Research Article

Medication use and risk of amyotrophic lateral sclerosis: using machine learning for an exposome-wide screen of a large clinical database

Ran S Rotem, Andrea Bellavia, Sabrina Paganoni & Marc G Weisskopf

Pages 367-375 | Received 24 Oct 2023, Accepted 12 Feb 2024, Published online: 01 Mar 2024

Cite this article https://doi.org/10.1080/21678421.2024.2320878 Check for updates

- High-dimensional screening of patients' history of medication use and ALS risk
- 501 ALS cases and 4,998 from a large health register
- 1000+ classes of medications (binary)

- Screening based on GBM
- Results: “8 consistently selected medications were more commonly used in cases compared to controls, with use prevalence in cases ranging from 1.4 to 30.3%”
- Selected classes were included in final logistic regression model

Table 3. Odds Ratios[a] (OR) and 95% confidence intervals (CI) from conditional logistic regression for selected ATC groups.

ATC group code	Main anatomic / pharmacologic group	Pharmacologic / therapeutic subgroup	chemical substance	OR (95% CI)	P-value
m01ah02	Musculo-skeletal system	Anti-inflammatory and Antirheumatic Products, Non-Steroids	Rofecoxib	1.51 (1.20-1.91)	<0.01
m03bx01	Musculo-skeletal system	muscle relaxants	Baclofen	5.07 (2.65-9.68)	<0.01
a06ad10	Alimentary tract and metabolism	drugs for constipation	osmotically acting laxatives	1.98 (1.44-2.73)	<0.01
r05da20	Respiratory system	Cough suppressants	opium alkaloids and derivatives	1.32 (1.04-1.69)	0.02
a11db	Alimentary tract and metabolism	B vitamins	vitamin b1 in combination with vitamin b6 and/or vitamin b12	1.47 (1.14-1.89)	<0.01
j01ee01	Anti-infective for systemic use	Sulfonamides and trimethoprim	sulfamethoxazole and trimethoprim	1.33 (1.02-1.74)	0.04
d07ca01	Dermatologicals	Corticosteroids with antibiotics	Hydrocortisone and antibiotics	1.37 (1.04-1.79)	0.02
a10ba02	Alimentary tract and metabolism	Blood glucose lowering drugs	Metformin	0.79 (0.56-1.13)	0.20

Example 2. Ohanyan et al. 2021

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

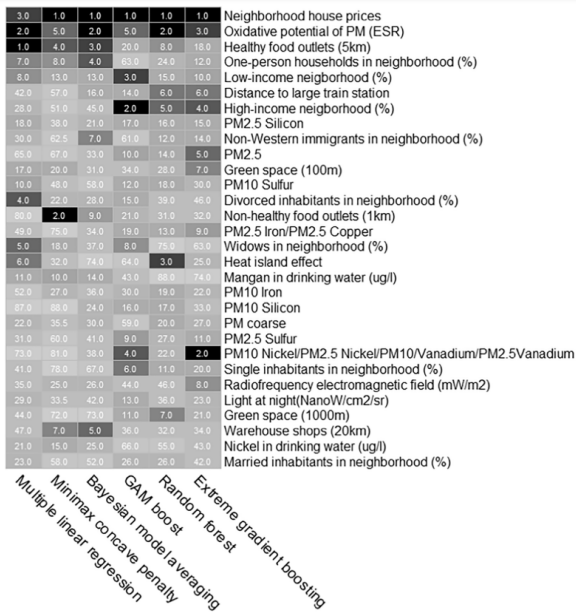
Machine learning approaches to characterize the obesogenic urban exposome

Haykanush Ohanyan [a,b,c,*], Lützen Portengen [b], Anke Huss [b], Eugenio Traini [b], Joline W. J. Beulens [a,c,d], Gerard Hoek [b], Jeroen Lakerveld [a,c], Roel Vermeulen [b]

- "To explore what environmental factors of the urban exposome are related to body mass index (BMI), and evaluate the consistency of findings across multiple statistical approaches" (Etiology)

- Data from 14,829 participants of the Occupational and Environmental Health Cohort study
- 86 environmental factors, including air pollution, traffic noise, green-space, built environmental and neighborhood sociodemographic characteristics
- Compared:[27]
  - Penalized Regression
  - Gradient boosting
  - xgboost
  - Multiple linear regression (for comparison)

[27] In addition to this list, two other potential methods often used in environmental epidemiology were also evaluated: Sparse Group Partial Least Squares, Bayesian Model Averaging


Multiple linear regression
Minimax concave penalty
Bayesian model averaging
GAM boost
Random forest
Extreme gradient boosting
Neighborhood house prices
Oxidative potential of PM (ESR)
Healthy food outlets (5km)
One-person households in neighborhood (%)
Low-income neigborhood (%)
Distance to large train station
High-income neigborhood (%)
PM2.5 Silicon
Non-Western immigrants in neighborhood (%)
PM2.5
Green space (100m)
PM10 Sulfur
Divorced inhabitants in neighborhood (%)
Non-healthy food outlets (1km)
PM2.5 Iron/PM2.5 Copper
Widows in neighborhood (%)
Heat island effect
Mangan in drinking water (ug/l)
PM10 Iron
PM10 Silicon
PM coarse
PM2.5 Sulfur
PM10 Nickel/PM2.5 Nickel/PM10/Vanadium/PM2.5Vanadium
Single inhabitants in neighborhood (%)
Radiofrequency electromagnetic field (mW/m2)
Light at night(NanoW/cm2/sr)
Green space (1000m)
Warehouse shops (20km)
Nickel in drinking water (ug/l)
Married inhabitants in neighborhood (%)
90
80
70
60
50
40
30
20
10
1

Additional applications of interest in exposome research

- Lampa et al. The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. Environmental health 2017
- Li et al. Two-step approach for assessing the health effects of environmental chemical mixtures: application to simulated datasets and real data from the Navajo birth cohort study. Environmental Health 2019
- Midya and Gennings. Detecting shape-based interactions among environmental chemicals using an ensemble of exposure-mixture regression and interpretable machine learning tools. Statistics in Biosciences 2024

For more details on specific applications in exposome research see chapter 5 of the mixture textbook

b) Clinical prediction modeling

Prediction in clinical epidemiology

- Integral part of medical care
- Predicting a future state of health based on information that is currently available (prognosis)
- Predicting a non-observable current state of health based on other available information (diagnosis)

Broadly speaking, we have 2 types of prediction:

- Prediction for binary outcomes: generate an estimate for the risk of developing outcome of interest (range 0 to 1)
- Classification: assigns subjects to risk categories (e.g. low risk, moderate risk, and high risk)

Examples [click for link]:

- MDcalc
- ASCVD 10-y risk
- TIMI calculators

Goals of an epidemiology study (etiology):

- Describe as accurately as possible an exposure-outcome association in the population of interest
- Investigate how the observed association describes a potential causal effect
- Often, hypothesis generation (e.g. identification of potential risk factors)
- -> We want to reduce bias and maximize precision

Goals of a prediction model:

- Use observed data to construct predictive tools that will be used on new individuals
- Be as pragmatic and parsimonious as possible
- -> We want to improve generalizability and applicability in different population/contexts
- Avoid overfitting

In general:

- When selecting potential confounders to control for in regression: be *liberal*
- When selecting potential variables to build a clinical prediction model: be *conservative*
- Nevertheless, the 2 worlds are overlapping
  - Epi studies will eventually require generalization
  - Prediction models target clinical decisions and often imply causal thinking [we are including risk factors, not associated factors]

This distinction has implications on the way we think of ML in epidemiology

- Etiology: screening and hypothesis generation
- Prediction: balance accuracy, overftting control, and ease of implementation

- Ongoing debate about the potential advantages of ML in clinical prediction
- Practicality and parsimony required by clinical risk scores might be achieved by regression and or LASSO
- Uncontrolled ML (e.g. feed all variables without pre-consideration) might add noise and challenges in the interpretability without considerable advantages
- Several papers have pointed out potential limitations
  - JCE systematic review
  - BMC simulation study
- Nevertheless, successful applications exist, especially when assessing the additional role of omics data on existing clinical scores. See, for example, Nurmohamed et al EHJ

- Some of these limitations might be partially due to the uncontrolled use
- Gradient boosting, for example, allows assessing the parsimony of the model with the shrinkage rate
- With ML, you decide what to feed to the machine, thus allowing human interventions for ethical decisions (e.g. including or not social determinants of health)

Example: Berg et al, in progress

- Research grant on proteomic discovery for heart failure risk in diabetic patients[28]
- $\sim$ 70 cases and 70 controls
- Blood sample used to quantify $\sim$ 400 proteins
- Validation of selected proteins in a larger cohort of 14,000 trial participants

Analytical challenge: how to identify the top predictors (out of $\sim$ 400) with only 70 events?

- Individual regressions are unreliable due to co-confounding and require multiple comparison assessment
- Multivariable models will not converge ($p > n$)
- Risk of overfitting

[28] Main results here

c) Other applications

Assessing heterogeneity of treatment effects over levels of multiple characteristics rather than one at the time

- ML can be used to assess effect modification over subgroups that are defined based on multiple patients' characteristics
  - Effect-based and risk-based stratification, using regression and random forest
  - Phenotype-based, using cluster analysis
- See an introduction in Bellavia & Murphy 2024 (link), and additional details in Hamaya et al 2025 (link)

Analysis of high-dimensional (e.g. omics) data

- Hypothesis generation and multivariable screening (eg proteomics discovery, identification of novel risk factors)
  - Rank through hundreds of potential covariates and use tools such as variables importance to identify potential factors that will be evaluated in future studies

4.6 Interpretability in etiologic research

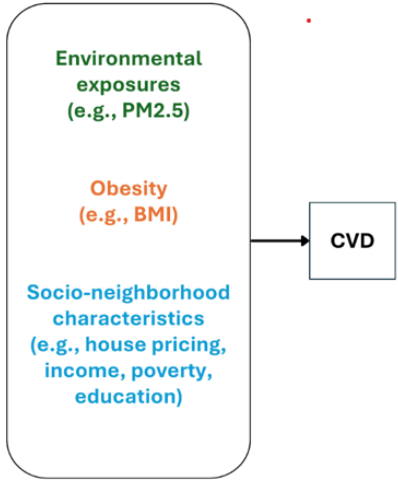
- Common methods for interpretable ML tackle the key yet preliminary aspect of statistical interpretability
- They provide tools for assessing and explaining the innner mechanisms (e.g. displaying functional forms without linearity assumptions, as splines do)

Causal interpretability

- This first set of interpretable tools scratches the surface and leaves several questions open
- Remember the goal of identifyinf actions to be taken
- The first question is whether the findings can be causally interpreted

Example

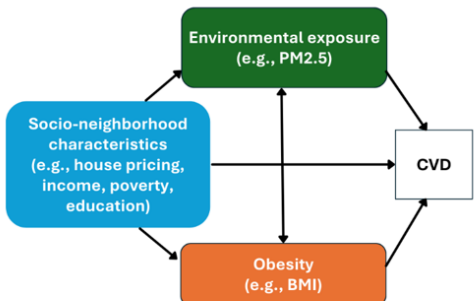
- Research study aimed at investigating the determinants of cardiovascular diseases (CVD) risk in the general population using an exposome-wide approach
- Hundreds of potential risk factors including pollutants and chemicals exposures, dietary factors, anthropometric measures, and neighborhood characteristics
- Interpretable ML was used to screen for potential risk factors


Environmental exposures (e.g., PM2.5)
Obesity (e.g., BMI)
Socio-neighborhood characteristics (e.g., house pricing, income, poverty, education)
CVD

- Suppose the ranking of top predictors included: 1) obesity, 2) neighborhood housing price, 3) PM2.5 levels, 4) SES
- Based on these results, can weconclude that BMI the main predictor of CVD?
- What about the association between PM2.5 and BMI?
- One could argue that BMI is a potential mediator of the PM2.5-CVD association
- If that was the case, we would not include BMI in a regression model. Same should happen if we use ML

- In more general terms, whenever you feed hundreds of variables to the machine you are implicitly making some causal assumptions
  - What if you include highly correlated variables?
  - What if you include both confounders and mediators?
  - What if you have repeated measurements over time?

- Nevertheless, there are simple steps we could take to improve causal interpretability of ML
- First, we could use high-dimensional DAGs to make hypothesis on the relationships between key variables



We could use these hypothesis to:

- Run separate ML models within each subcategory
- Consider methods for high-dimensional mediation
- Operate variable selections to screen among highly correlated predictors

Key message: when swithing to ML for etiological research, do not forget about causality! The machine does not distinguish confounders/mediators/main predictors and will treat everything equally

Causal ML

- Causality in ML is a very active area of research. Potential approaches include:[29]
  - Targeted Maximum Likelihood Estimation (TMLE)
  - Augmented Inverse Probability Weighting (AIPW)
  - Double/Debiased Machine Learning (DML)

[29] See the excellent review of Moccia et al.

Actionable interpretability

- Third level of interepretability required for identifying potential actions
- Say we have indeed identified a set of pathways and we are confident these represent causal pathways
- Next key questions:
  - Are these causal pathways actionable?
  - Are these actions going to be effective for all individuals in the population?

Social determinants of health (SDH)

- In our illustrative example, SES was identified as a top predictor
- This provides limited room for practical intervention
- SES, like other SDH, encompasses multiple dimensions of social and economic factors that are often hard to individually quantify
- Nevertheless, SDHs are integral to exposomic research and should be integrated. The key quesitons is how
- Relevant topic in epi research in general, even outside of ML applications

- First, unpack the meaning of the social factors and try to identify real determinants of health and actionable components
- For example, assessing more detailed information on proximal determinants of SES such as income, housing type, or education, allows identifying more clear and actionable routes of interventions and recommendations
- Ideally, this should happen at the study design phase for collecting required data

A key component of ML applications in epidemiology is a careful human-driven selection of what variables should be fed to the machine. This should be informed by causal assumptions, ethical and practical considerations, and driven by the goal of identiying actionable interventions and public health recommendations.

Assessing effect modification with complex data

- Final question, are the identified actions effective for all individuals?
- In regression, we would address this question with interactions and stratified analyses
- Similar approaches can be applied when using ML in complex settings
- A simple preliminary step is to run different models within population substrata

- More formal applications of ML for stratifying over multiple characteristics simultaneously have been discussed in clinical epidemiology for the assessment of multivariable heterogeneity of treatment effects (HTE)[30]

[30] See Bellavia/Murphy 2025, Circulation (link here) for a general overview

5. Additional topics

5.1 Unsupervised ML

Techniques of data reduction/classification based on a set of features (i.e. exposures/predictors) irrespective of specific labels (i.e. outcome). These include:

- Principal Component Analysis (dimensionality reduction from p covariates to less than p principal components)
- Cluster Analysis (individual groupings based on shared characteristics)

Cluster analysis

- Increasing popularity in clinical research for patients' phenotyping. Examples:
  - Cardiogenic shock[31]
  - Acute illness[32]
  - Sarcoidosis[33]

[31] https://www.jacc.org/doi/10.1016/j.jacadv.2022.100126
[32] https://www.nature.com/articles/s41598-024-59047-x
[33] https://respiratory-research.biomedcentral.com/articles/10.1186/s12931-022-01993-z

- Cluster analysis allows classifying individuals based on their exposure profiles
- The main approaches for cluster analysis are:[34]
  - Hierarchical clustering
  - K-means clustering
  - Model-based clustering
  - Density-based clustering
  - Fuzzy clustering

[34] See this webpage for a general overview of these approaches

K-means clustering

Classify individuals in k groups so that individuals within the same cluster are as similar as possible, while individuals from different clusters are as dissimilar as possible.

1. Pre-specify k
2. Select k random individuals as center for each cluster and define the centroids, vectors of length p with the means of all variables for the observation in the cluster
3. Define a distance measure. The standard choice is the Euclidean distance defined as $(x_i - \mu_k)$
4. Assign each individual to the closest centroid
5. Update the cluster centroid by calculating the new mean values of all the data points in the cluster
6. Iteratively update the previous 2 steps until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations

Model-based clustering

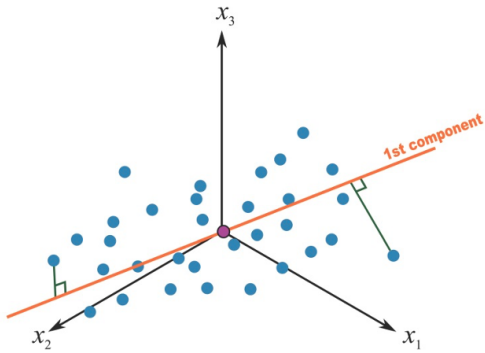
- Improves over K-means as it also allows incorporating continuous predictors
- Plus, additional features such as potentially operating variable selection
- Comprehensive and well documented R package VarSelLCM[35]

[35]We also have R code publicly available on Github (here) for clustering-based HTE assessment in clinical trials

PCA

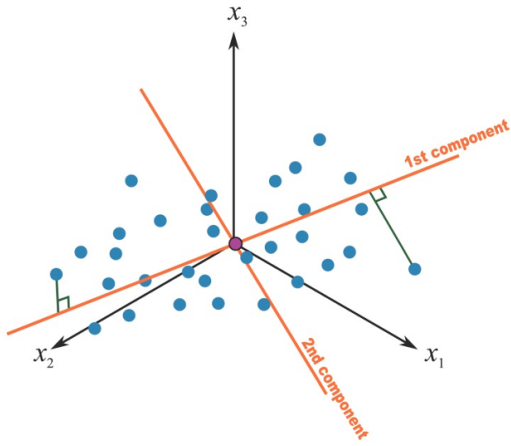
- PCA allows a better visualization of the variability present in a dataset with many covariates
- This "better visualization" is achieved by transforming a set of covariates into a smaller set of principal components
- Given a set of covariates, a first principal component is geometrically identified as the he straight line that best spreads the data out when it is projected along it, thus explaining the most substantial variance in the data

Example with 3 covariates: first component



Partial dependence plots

- Mathematically, this first component t_1 is calculated as a linear combination of the p $T = XW_p$, where W_p are the weights that maximize the overall explained variability
- These weights correspond to the eigenvectors of the correlation matrix
- A second component is than calculated by maximizing the residual variance, and under the constraint of orthogonality (uncorrelation) with the first one



Partial dependence plots

- We then proceed by calculating p components out of the original p covariates
- Practically speaking, we are transforming a set of p correlated variables into a set of p uncorrelated principal components
- PCA is sensitive to unscaled covariates, so it is usually recommended to standardize your matrix of exposures
- Deriving the p components is based on mathematical rules; the next steps of a PCA will instead be mostly subjective.

How many components?

- p components are required to explain 100% of the original variability. Our original goal, however, was dimensionality reduction.

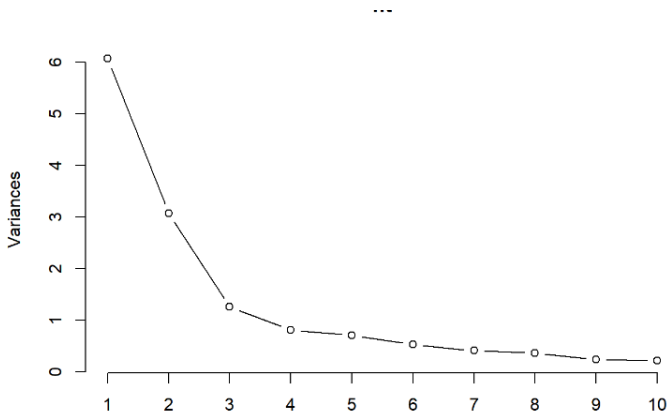
- How many components should be retrieved is subjective, but several tools are available to aid the decision:
  - Select components that explain at least 70 to 80% of the original variance
  - Select components corresponding to eigenvalues larger than 1
  - Look at the point of inflation in the scree plot
  - Consider components interpretation

R output of a PCA on 14 covariates

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     2.4627 1.7521 1.12071 0.89784
## Proportion of Variance 0.4332 0.2193 0.08971 0.05758
## Cumulative Proportion  0.4332 0.6525 0.74219 0.79977
##                            PC5     PC6     PC7     PC8
## Standard deviation     0.83905 0.72337 0.63861 0.60268
## Proportion of Variance 0.05029 0.03738 0.02913 0.02594
## Cumulative Proportion  0.85006 0.88744 0.91657 0.94251
##                           PC9    PC10    PC11    PC12
## Standard deviation     0.4892 0.46054 0.43573 0.29751
## Proportion of Variance 0.0171 0.01515 0.01356 0.00632
## Cumulative Proportion  0.9596 0.97476 0.98832 0.99464
##                           PC13    PC14
## Standard deviation     0.25542 0.09904
## Proportion of Variance 0.00466 0.00070
## Cumulative Proportion  0.99930 1.00000
```

Partial dependence plots

Scree plot:



Partial dependence plots

Components interpretation

- To interpret components we need to look at the loading factors, the correlation coefficients between the derived components and the original covariates
- It is recommended to perform axes rotation (e.g. varimax, handled by all R packages) for optimal interpretation.

Example from Sanchez et al. 2018

Standardized rotated factor loading and communalities for specific gravity standardized metals (n = 199).

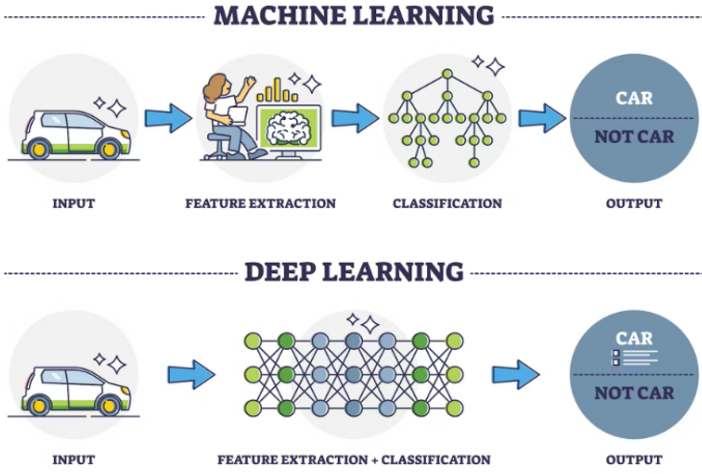
Component	1	2	3	4	5	6
As	−0.16	0.12	**0.63**	0.09	−0.16	0.22
Ba	0.16	0.08	−0.08	**0.61**	0.10	−0.20
Cd	0.02	−0.05	−0.01	0.03	0.15	**0.64**
Co	0.20	0.26	0.06	−0.02	−0.23	0.36
Cs	**0.52**	−0.03	−0.05	−0.01	−0.13	0.21
Cu	0.10	−0.04	0.01	0.09	**0.46**	0.28
Mn	−0.08	−0.06	0.07	**0.75**	−0.06	0.11
Mo	0.27	−0.06	**0.44**	−0.15	0.05	−0.08
Ni	**0.55**	−0.10	−0.01	−0.05	0.03	0.06
Pb	−0.03	**0.60**	0.06	−0.03	0.15	−0.24
Se	0.09	0.38	−0.02	−0.04	0.18	0.06
Sr	**0.48**	0.12	−0.03	0.10	−0.01	−0.31
Tl	−0.04	**0.59**	−0.05	0.01	−0.12	0.15
W	0.05	−0.09	**0.62**	0.01	0.14	−0.19
Zn	−0.05	0.04	−0.01	−0.03	**0.76**	0.02
Eigenvalue	3.17	2.33	1.88	1.61	1.50	1.49
Total variance	21.11%	15.55%	12.51%	10.72%	10.27%	9.98%
Cumulative variance	21.11%	36.65%	49.17%	59.89%	70.16%	80.14%

Factor loadings are bolded if > 0.40.

5.2 Introduction to Deep learning

- Machine learning still requires human interventions at different stages. For example:
  - Feature extraction: the starting data is a set of rows and columns like the one we use for regression. This is often made by humans (e.g. manually filling a clinical record). In technical terms, we are analysis tabular data
  - Re-tuning: if the machine makes an error, we can manually work on some parameters tuning to potentially improve the performance
- Deep learning provides considerable advantages to those settings where the data is too complex for even these 2 tasks to be left to the human

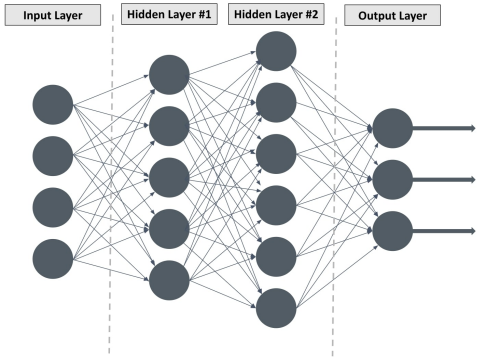
- An obvious setting is when the big data is really big. If we deal with millions of rows/columns we greatly benefit from a machine that can correct itself as there are too many parameters' combinations we might consider
- The other major setting, with considerable clinical applications, is pattern recognition
- This has provided incredible advances to the field of Medical Image Analysis
  - Identify abnormalities in large-scale pathology images
  - Brain tumor segmentation in MRI
  - Analyze echocardiograms, MRI, CT scans for heart disease diagnosis
  - Detect fractures in X-rays
  - ...



Common representation of ML vs DL

Neural networks

- Deep learning algorithms literally try to imitate the way the human brain works
- They do so by using a layered structure of algorithms called artifical neural networks (ANN)
- These include an input layer, a set of hidden layers, and an output layer



General notes on ANN

- ANN can be used for both supervised and unsupervised analysis
- They perform at their best with large volumes of data. With small data they usually provide inaccurate results and high variance -> ML should be preferred in these settings
- They require extensive computational abilities
- Advantages/limitations in epi studies including clinical prediction modeling still under debate. See for example Unterhuber et al 2021