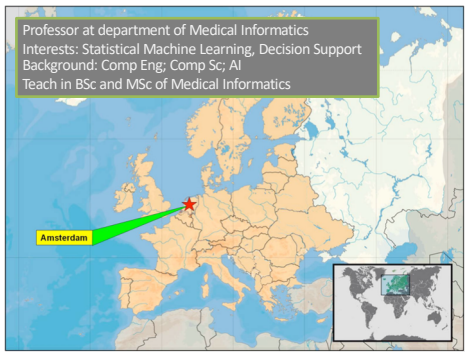
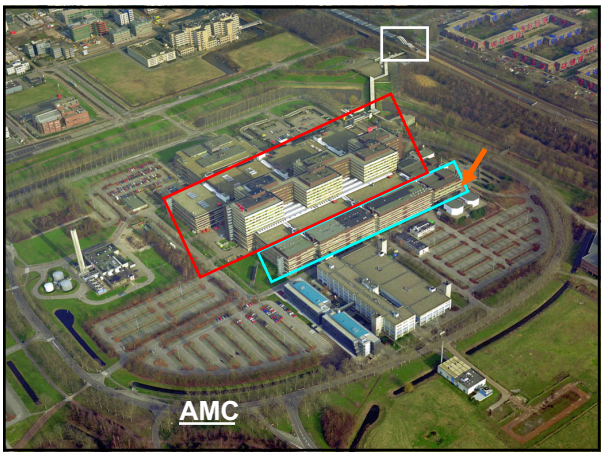
Tutorial @ AIME 2019

Evaluating Prediction models in Medicine

Ameen Abu-Hanna
Dpt. Medical Informatics
University of Amsterdam
(a.abu-hanna@amsterdamumc.nl)

About me





Overview

1. Introduction
2. Performance measures
3. Model selection and validation
4. Challenges
5. Seven steps to develop and validate your model

Various slides have been co-created with Niels Peek.

Overview

1. Introduction
2. Performance measures
3. Model selection and validation
4. Challenges

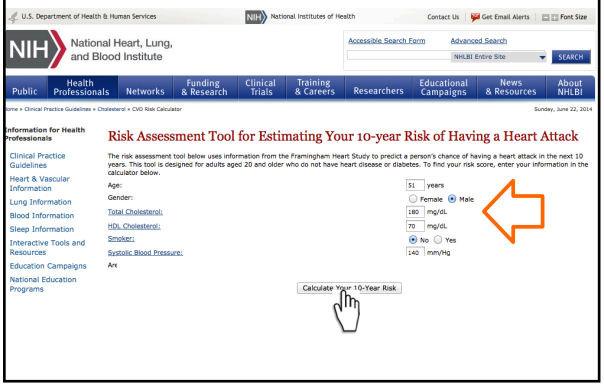
Medical prediction models

Provide risk estimates for:

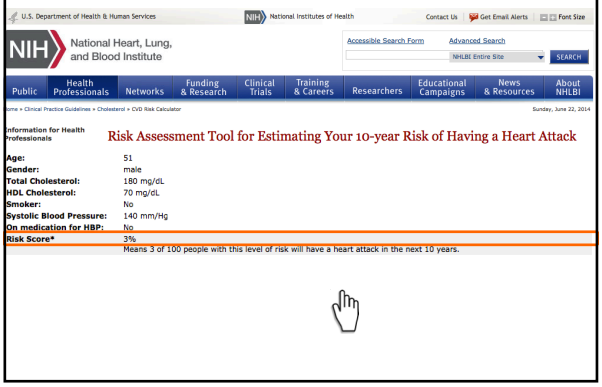
- Diagnosis: Presence of disease
- Prognosis: Future course and outcome of disease
  - This may concern the natural history of disease or the results of treatment

Driver: Personalized medicine

Example: Cardiac risk calculator



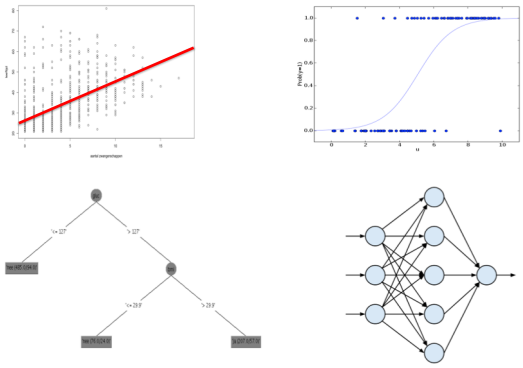
Example: Cardiac risk calculator

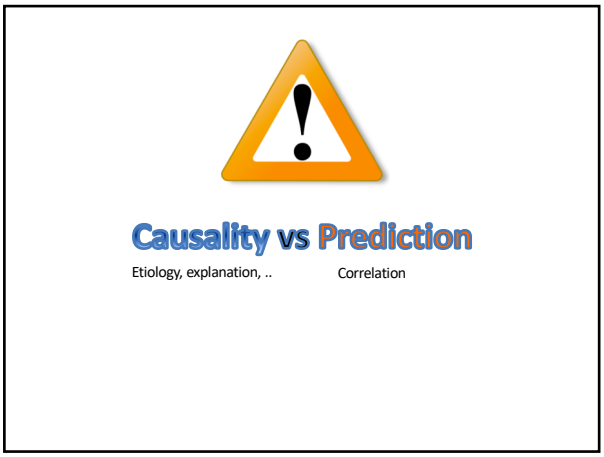


#preg	gluc	bldprs	...
6	148	72	
1	85	66	
8	183	64	
1	89	66	
0	137	40	
5	116	74	
3	78	50	
10	115	0	
2	197	70	
8	125	96	
4	110	92	
10	168	74	
10	139	80	
1	189	60	
5	166	72	
7	100	0	
0	118	84	
7	107	74	
1	103	30	
1	115	70	
3	126	88	
	. . .		

Learning Algorithm

- Linear regression
- Logistic regression
- Cox regression
- Decision trees
- Support Vector Machines
- Neural Networks
- ...





Causality vs Prediction

Etiology, explanation, .. Correlation

Motivating Example

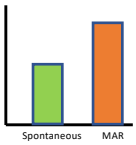
Risk of poor neonatal outcome at term after medically assisted reproduction: a propensity score–matched study

Sabine Ensing, M.D.,[a,b] Ameen Abu-Hanna, Ph.D.,[b] Tessa J. Roseboom, Ph.D.,[b,c] Sjoerd Repping, Ph.D.,[d] Fulco van der Veen, M.D., Ph.D.,[d] Ben Willem J. Mol, M.D., Ph.D.,[a,e] and Anita C. J. Ravelli, Ph.D.[b]

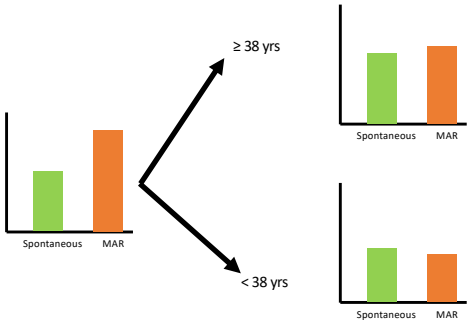
Fertility and Sterility® 2015

Problem

- 1,953,932 pregnancies from a national registry for 1999–2011
- 48,921 (2.5%) were pregnant after MAR (Medically Assisted Reproduction)
- Apgar score worse in MAR: 5.3% vs 3.4%
  - OR 1.59 (1.28–1.96).
- Is MAR the culprit?



Stratification (hypothetical situation)



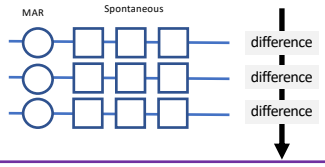
Age, and many other factors, could be confounding the relationship.
We must account for this confounding.

Possible confounders

- maternal age
- ethnicity
- socioeconomic status
- parity
- year of birth
- preexistent diseases

Hard to stratify on all of these confounders. We will match each MAR delivery with very similar spontaneous deliveries in terms of these confounders.

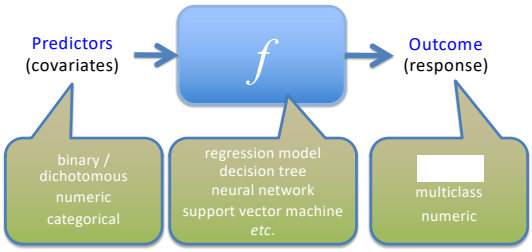
Adjusting for confounders by matching



- Propensity score matching: No difference.
- OR 0.99 (0.87–1.14).

Conclusion: MAR is a predictor but not a cause.

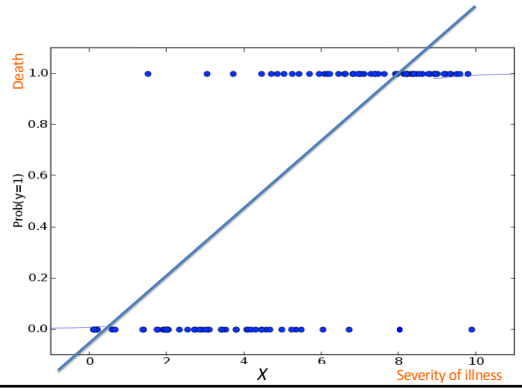
Prediction model



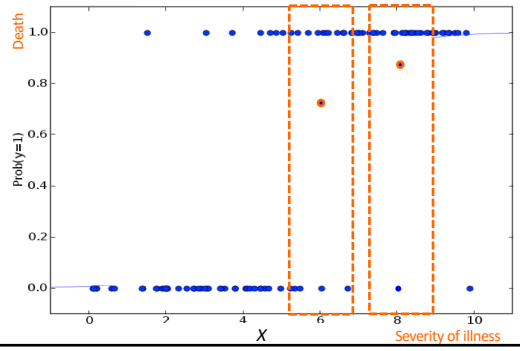
Probabilistic Classification

- Prediction applications often require models that predict probabilities rather than classes
- This is called probabilistic classification
- Most classification methods (e.g. decision trees) can be adapted for probabilistic classification
- Logistic regression is a statistical method that specifically aims at probabilistic classification

Linear regression? No

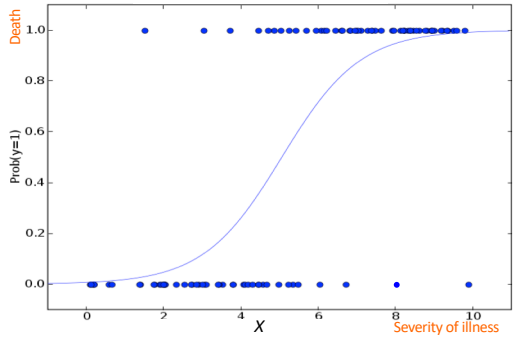


Non-parametric



Logistic regression

$$\text{logistic}(x) = \frac{1}{1+e^{-x}}$$



In general: Linear predictor

$$LP = \beta_0 + \beta_1 age + \beta_2 sex + \ldots$$

$$\text{logistic}(LP) = \frac{1}{1+e^{-LP}}$$

Why to predict?

Applications of prediction models

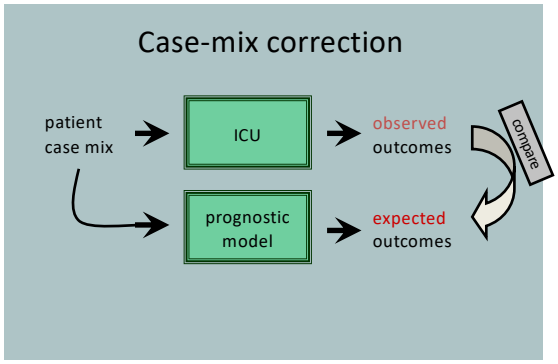
- **Public health**
  - identify high risk for preventive actions
- **Clinical practice**
  - risk of having disease (start intervention?)
  - informing patients/families on risk (withdraw therapy?)
  - probability of outcome w. various treatments (treatment selection)
- **Research**
  - inclusion in clinical trials (e.g. only high risk patients)
  - case-mix risk adjustment (benchmarking centers) [→]

Benchmarking ICUs

Suppose we wish to compare the performance of the intensive care units (ICUs) in two different hospitals.

	Poznan	Amsterdam
		
Mortality:	25%	15%

Case-mix correction



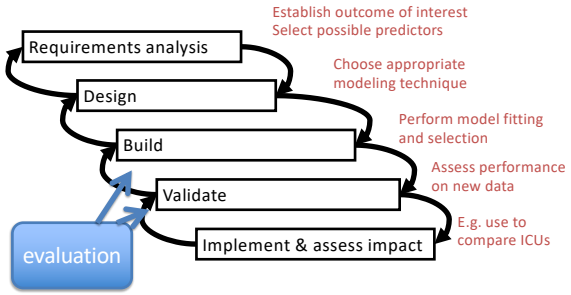
Benchmarking ICUs

	A	B
		
Observed mortality:	25%	15%
Expected mortality:	30%	12%

Estimates from prognostic model

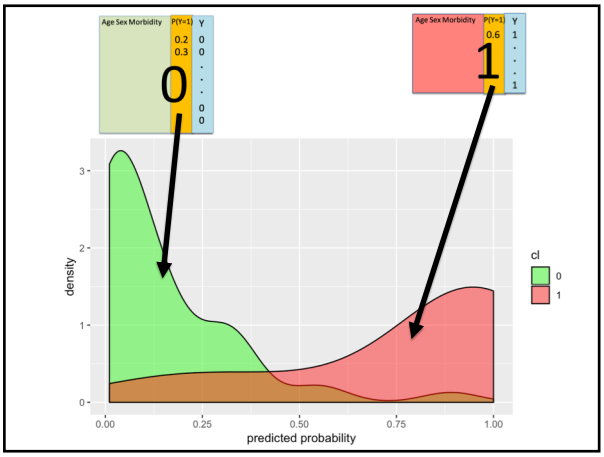
Standardized Mortality Ratio:	0.83	1.25

Prediction model life-cycle



Overview

1. Introduction
2. Performance measures
3. Model selection and validation
4. Challenges



Two fundamental concepts

Discrimination

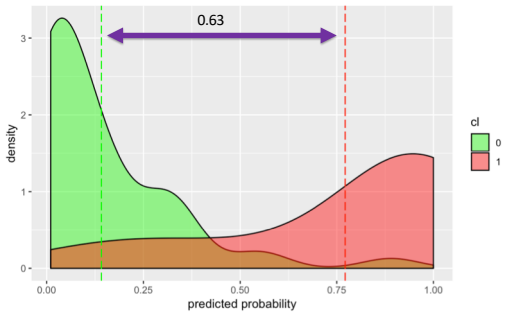
- does model discriminate between positive and negative outcomes?
- in general, positive cases should have a higher probability than negative cases

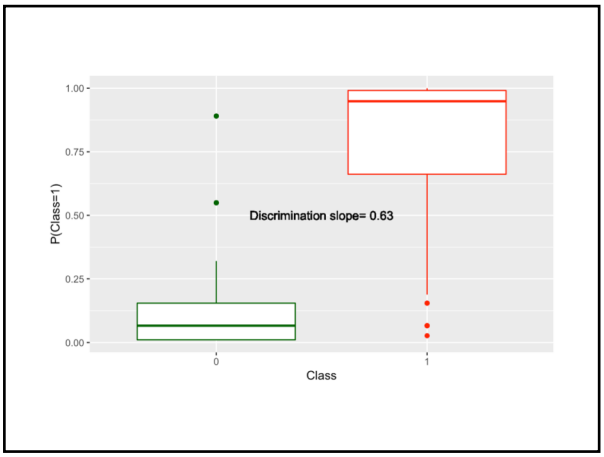
Calibration

- are the probabilities predicted by the model close to the real probabilities?
- more difficult to study because we don't know the real probabilities

Discrimination

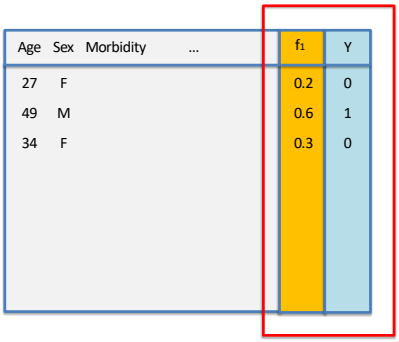
Discrimination Slope
Area under the AUC





Area Under the ROC Curve (AUC)

Instead of grouping in class = 0 and class = 1 we consider the P and Y for each patient



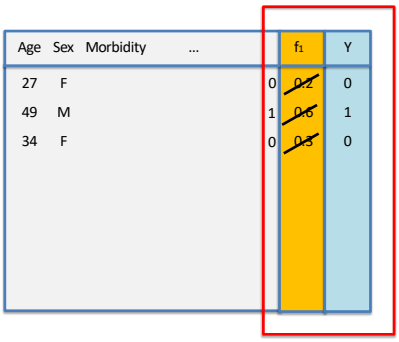
Age	Sex	Morbidity	...	f_1	Y
27	F			0.2	0
49	M			0.6	1
34	F			0.3	0

Discritization-based measures

A battery of accuracy measures is based on discretization of f_1 into 0 and 1 to obtain $\hat{Y}$, before comparing to Y, e.g.

$$\hat{Y} = \begin{cases} 1 & \text{if } f_1 > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Discretizing f_1 with threshold = 0.5



Age	Sex	Morbidity	...	f_1	Y
27	F			0 ~~0.2~~	0
49	M			1 ~~0.6~~	1
34	F			0 ~~0.3~~	0

Confusion matrix

Most discrimination-based measures can be defined based on the confusion matrix:

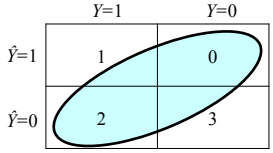
Prediction \ Outcome	$Y=1$	$Y=0$
$\hat{Y}=1$	True Positives	False Positives
$\hat{Y}=0$	False Negatives	True Negatives

Example

Y	$\hat{Y}$
0	0
1	0
0	0
1	0
0	0
1	1

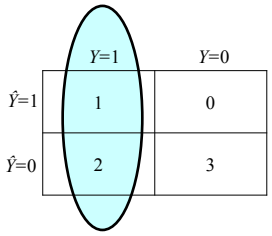
	$Y=1$	$Y=0$
$\hat{Y}=1$	1	0
$\hat{Y}=0$	2	3

Error rate



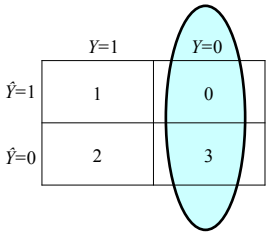
- $\frac{(FP + FN)}{(TP + TN + FP + FN)}$
  = (0+2)/ (1+3+0+2)
  = 2/6 = 1/3
- Proportion of incorrect classifications: 2 out of 6 instances have been incorrectly classified

Sensitivity ("Recall" in Inf Retrieval)



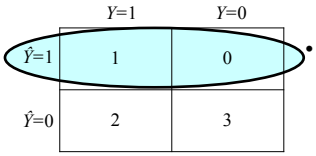
- TP / (TP + FN) = 1 / (1 + 2) = 1/3
- How "sensitive" is the prediction to the **event**: only 1 out of the 3 events has been "detected"
- Also called hit rate

Specificity



- TN/(TN + FP) = 3 /(0 + 3) = 1
- Like sensitivity but for the non-event: all three non-events have been "detected"
- 1-specificity is a.k.a. false alarm rate

Positive predictive value
("Precision" in Inf Retrieval)



- TP/(TP + FP) = 1/(1 + 0) = 1
- From those of which the model predicted the event, all indeed had the event

ROC analysis

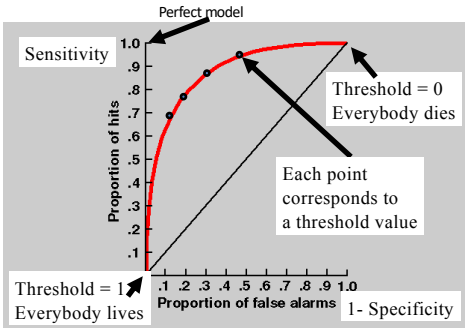
- A cut-off point has 1 confusion matrix
- To get an aggregate performance measure we make use of many cut-off points between 0 and 1
- For each we calculate the sensitivity and specificity
- Plotting the sensitivity versus 1-specificity we get a ROC curve

Ŷ for various thresholds

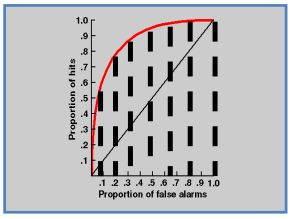
For each $\hat{Y}$ column calculate Sensitivity and Specificity

Y	f_1	$\hat{Y}(0)$	$\hat{Y}(0.15)$	...	$\hat{Y}(0.5)$	...	$\hat{Y}(1)$
0	~~0.1~~	1	0	...	0	...	0
0	~~0.25~~	1	1		0		0
...				...		...	
1	~~0.3~~	1	1		0		0
				...		...	
1	~~0.9~~	1	1		1		0

Example ROC curve



Area under ROC curve (AUC)



If you randomly choose two instances:

d from those who died

a from those alive

then the AUC is the proportion of cases in which **d** has a higher probability than **a**

Example

000000000000000000000000 1111 00000 111111

25 4 5 6

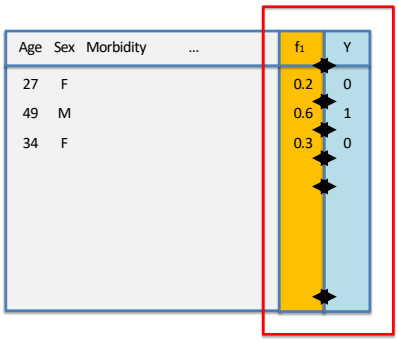
What is the AUC?

AUC = (25*4 + 30*6)/300 = 0.93

Characteristics of measures for discrimination

Measure	Calculation	Visualization	Pros	Cons
Concordance statistic	Rank order statistic	ROC curve	Insensitive to outcome incidence; interpretable for pairs of patients with and without the outcome	Interpretation artificial
Discrimination slope	Difference in mean of predictions between outcomes	Box plot	Easy interpretation, nice visualization	Depends on the incidence of the outcome

General (in)accuracy measures



General (in)accuracy measures

- Mean squared error (Brier score)

$$\frac{\sum_1^N (f_{1,i} - y_i)^2}{N}$$

Ex1. For $i = 7$ (patient #7) if $f_{1,7}$ = 0.3 and $y_7 = 1$, then the Brier Score is $(0.3-1)^2$ = 0.49.

Ex2. For $i = 12$ (patient #12) if $f_{1,12}$ = 0.1 and $y_{12} = 0$, then the Brier Score is $(0.1-0)^2$ = 0.01.

Observation

- Low Brier score values indicate better accuracy
- However, the worst (highest) value depends on the prediction problem at hand
- A normalized version of the Brier score is called R^2 (or the Brier Skill Score).

R^2 or Brier Skill Score

Usually defined as

$$R^2 = 1 - \frac{\sum_1^N (f_{1,i} - y_i)^2}{\sum_1^N (y_i - \bar{y})^2}$$

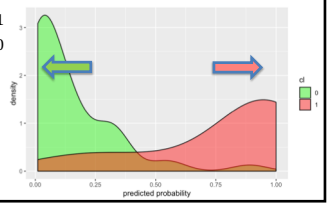
where $\bar{y}$ is the average value of the y_i s

The denominator corresponds to the error of a nondiscriminative model (variance in this case).

Net Reclassification Improvement (NRI) (for comparing two models)

- Determine the extent to which the new predictive model **improves** the classification
- Here we use the "continuous" version (Pencina Stat Med 2008)
  - NRI = # times new predictor improves upon old predictor less the # times the new predictor is inferior.
  - An **improvement** is defined by

if $f_1^{new} > f_1^{old}$ for $Y = 1$
if $f_1^{new} < f_1^{old}$ for $Y = 0$



Peculiar behavior?

- Suppose your predictions are the true probabilities of the outcome
- Calculate the Area under the ROC curve
- Now square each of the predictions
- Re-calculate the AUC for these (bad) predictions
- then *the AUC … does not change*!

Important properties

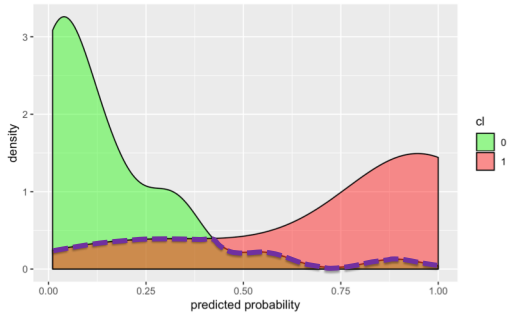
- Non-Strictly proper measure:
  Minimum error is achieved when prediction = true probability but *also with other predictions*
- Strictly proper measure:
  Minimum error obtained *only when* prediction = true probability of outcome

Observations

- Non-strictly proper
  - Error rate, AUC, NRI
- Strictly proper
  - Brier score, R^2
- AUC, NRI, and R^2 have intuitive meaning

Imbalanced datasets

- When prevalence is low the AUC can mislead us
- A high AUC does not yet tell us that the PPV ("precision" in IR) is high, which is more relevant in these cases
- The PPV decreases when the prevalence is low



The larger the tail of Class 0 (Survival) is in relation to Class 1 (Death) the less the PPV is, but the AUC is not much affected.

Example 1

A B C D E F G H I J K L

000000000000000000000000 1111 00000 111111

25 4 5 6

A. PPV = 10/15 sens = 10/10
B. PPV = 9/14 sens = 9/10
C. PPV = 8/13 sens = 8/10
D. PPV = 7/12 sens = 7/10
E. PPV = 6/11 sens = 6/10
F. PPV = 6/9 sens = 6/10
G. PPV = 6/6 sens = 6/10
H. PPV = 5/5 sens = 5/10
I. PPV = 4/4 sens = 4/10
J. PPV = 3/3 sens = 3/10
K. PPV = 2/2 sens = 2/10
L. PPV = 1/1 sens = 1/10

Example 2 increasing negatives

A B C D E F G H I J K L

000 … 000 1111 000 … 000 111111

2500 4 500 6

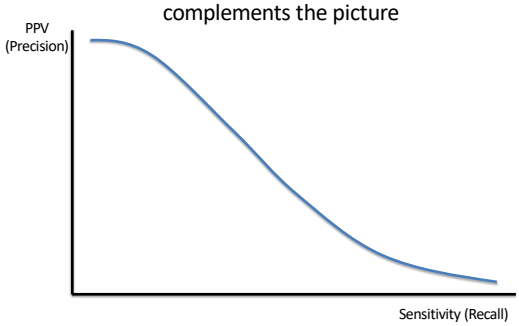
A. PPV = 10/510 sens = 10/10
B. PPV = 9/509 sens = 9/10
C. PPV = 8/508 sens = 8/10
D. PPV = 7/507 sens = 7/10
E. PPV = 6/506 sens = 6/10
F. PPV = 6/256 sens = 6/10
G. PPV = 6/6 sens = 6/10
H. PPV = 5/5 sens = 5/10
I. PPV = 4/4 sens = 4/10
J. PPV = 3/3 sens = 3/10
K. PPV = 2/2 sens = 2/10
L. PPV = 1/1 sens = 1/10

From previous slide

A. PPV = 10/15 sens = 10/10
B. PPV = 9/14 sens = 9/10
C. PPV = 8/13 sens = 8/10
D. PPV = 7/12 sens = 7/10
E. PPV = 6/11 sens = 6/10
F. PPV = 6/9 sens = 6/10
G. PPV = 6/6 sens = 6/10
H. PPV = 5/5 sens = 5/10
I. PPV = 4/4 sens = 4/10
J. PPV = 3/3 sens = 3/10
K. PPV = 2/2 sens = 2/10
L. PPV = 1/1 sens = 1/10

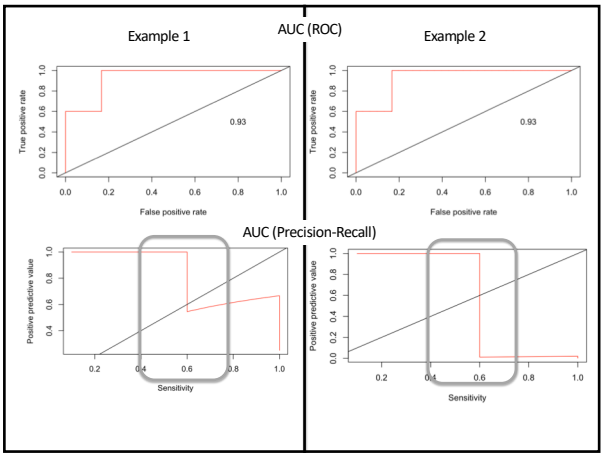
Precision-Recall* Plot

complements the picture



*Precision and recall as used in information retrieval

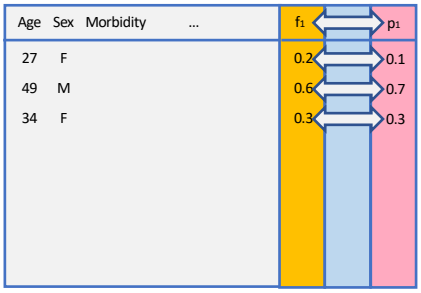
Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, 1–21 (2015).



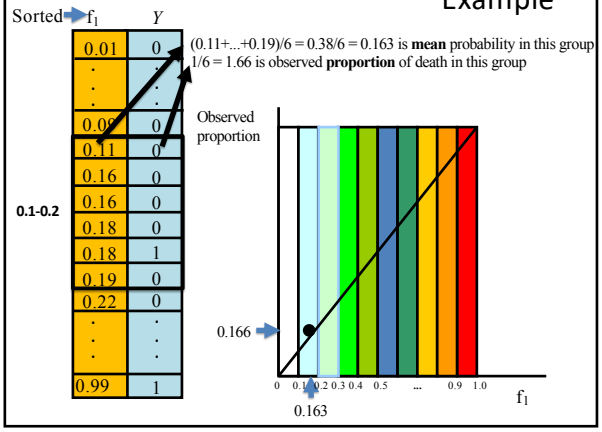
Calibration

We call this also precision but this is different than precision in information retrieval so we will stick to calibration

Calibration

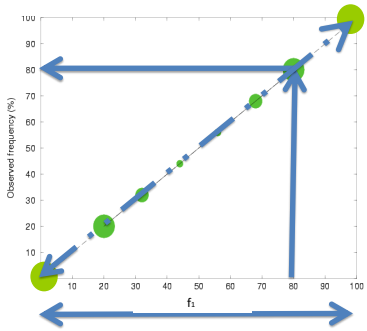


Example



Calibration in context with other aspects

Ideal calibration and range, very good sharpness



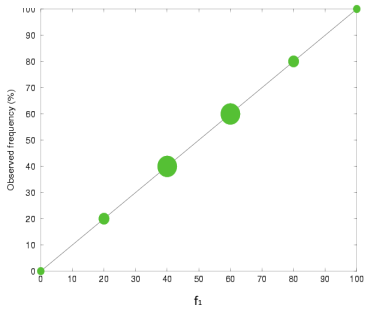
Measure of (un)sharpness

unsharpness = $\frac{1}{N}\sum_1^N f_{1,i}(1-f_{1,i})$

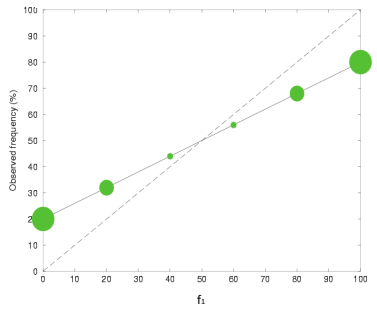
Maximum sharpness when unsharpness=0 (when all predictions are 0 or 1).

Lowest sharpness when all predictions = 0.5 yielding unsharpness = 0.25

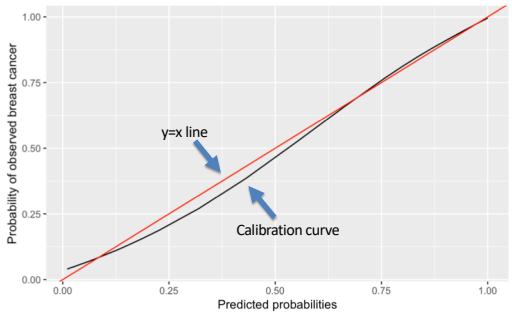
Good calibration, but poor sharpness



Poor calibration, but good sharpness



Realistic calibration plot obtained by smoothing



Valuable, or merely valid?

Model validation involves more than statistical evaluation of a model's performance.

Often there are additional context-dependent criteria that determine a model's use and value

Altman DG, Royston P. *Statistics in Medicine* 2000; **19**:453–73.

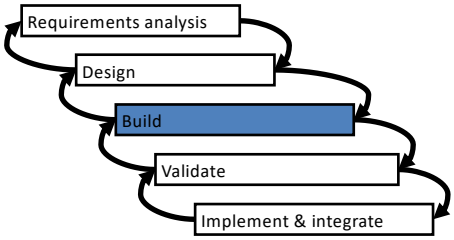
Factors affecting use

- Qualtiy
  - Large, high quality dataset
  - Based on a study protocol with a sound statistical analysis plan
  - Validated in independent datasets from different locations
- Usefulness
  - Timely, readily available predictors, causal
  - Intuitive and unambiguous predictors
  - Model complexity and format. Ease of use in consulting room.
  - Decision making guidance (interpretation of probs).
- Endorsement by leading professionals

Overview

1. Introduction
2. Performance measures
3. Model selection and validation
4. Challenges

Model selection



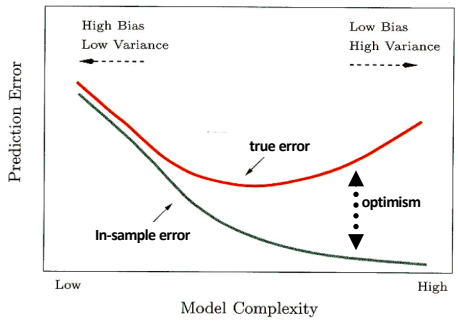
The model selection problem

During the model building phase, we can often choose from a number of models that are related in structure but vary in complexity.

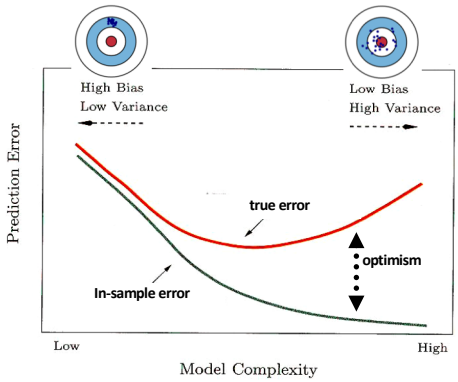
Examples:
- small decision trees vs. large decision trees
- number of nodes in a neural network
- selection of covariates in regression models

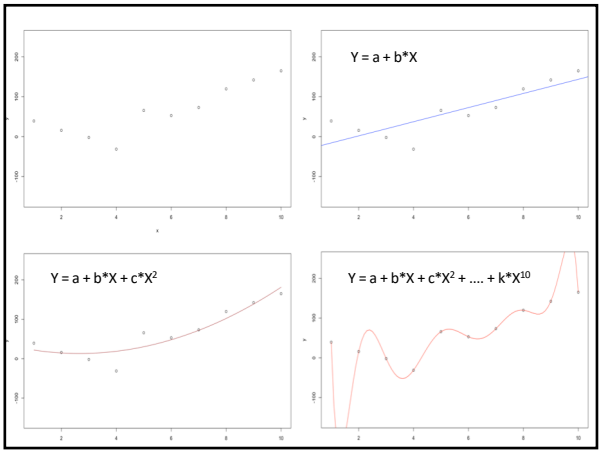
Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001, Chapter 7.

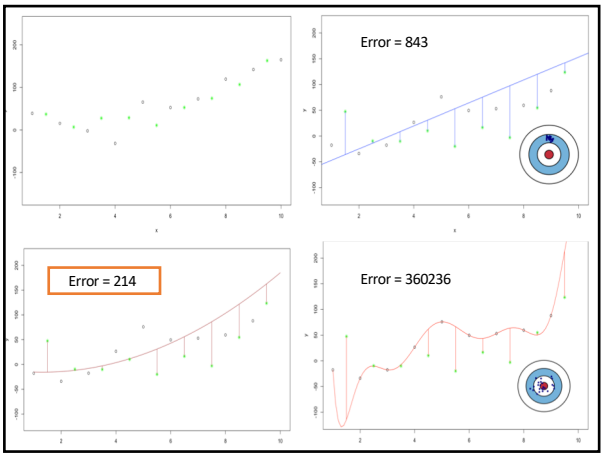
The bias-variance trade-off



The bias-variance trade-off



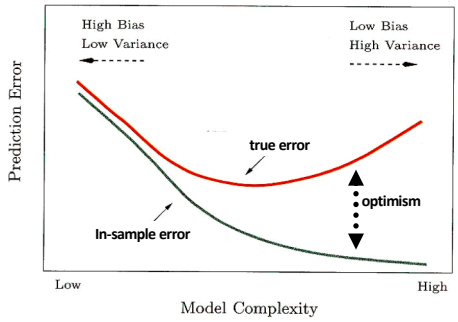




The bias-variance trade-off

Parametric models (linear, logistic, Cox regression)

Non parametric models (Decision trees, Neural networks)



Two lessons

- We cannot trust performance on the training set
- We need to control for model complexity

The split-sample procedure for validation

training set	test set

1. Data set is randomly split into training set and test set (usually 2/3 vs. 1/3)
2. Fit model on training set
3. Measure performance on test set

Cross-validation

fold 1	fold 2	…	fold *k*

1. Split data set randomly into *k* subsets ("folds")
2. Build model on *k*-1 folds
3. Compute error on remaining fold
4. Repeat *k* times

- Average error on *k* test folds approximates true error on independent data, of the model that is built on all *k* folds

How to control complexity of model?

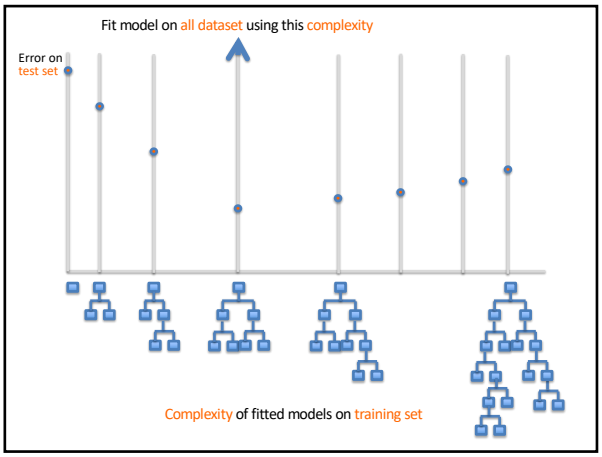
- Use independent observations
  - split sample
  - cross-validation
- Adjust for model complexity
  - AIC
  - Shrinkage
  - Lasso

Hand DJ. Construction and Assessment of Classification Rules, John Wiley & Sons, 1997, Chapter 7.

The split-sample procedure for selecting model complexity

training set	test set

1. Data set is randomly split into training set and test set
2. Model with various complexities are built on training set and predictive performance is measured on test set
3. Select optimal complexity
4. Fit model with optimal complexity on all data



What is expected performance?

But wait.. we used all the data to fit the final model so you should not report the performance you obtained with the optimal complexity.

What can you do to estimate performance?

Use a completely fresh untouched dataset to test on (use part of the training set itself to find optimal performance and test on test set).

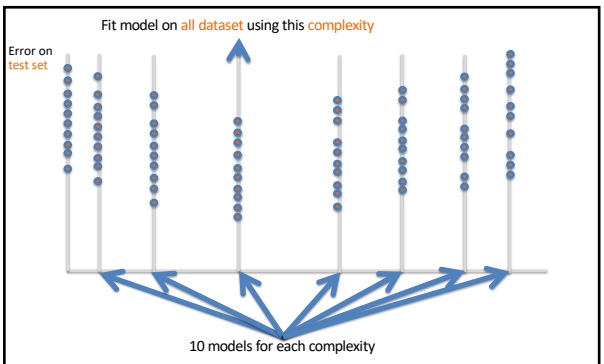
You will still report the final model but now you can provide an honest estimate of its expected performance on unseen data.

Cross-validation for selecting model complexity

fold 1	fold 2	...	fold *k*

1. Split data set randomly into *k* subsets ("folds")
2. Build models with various complexities on *k*-1 folds
3. Compute average performance of these models on remaining fold
4. Repeat *k* times
5. Select optimal complexity
6. Fit model with optimal complexity on all dataset

Again: we use the procedure above to decide on complexity and best model. We need to use a fresh dataset to have honest estimate of performance. We can also do a nested-cross validation.



What is expected performance?

You can use a nested cross validation.

The "inner loop" is used to adjust complexity

The "outer loop" is used to measure performance

Adjust for model complexity
Information criteria

- AIC = Akaike Information Criterion
- Formula:

  $$AIC = -2 \cdot \text{loglik} + 2 \cdot d$$

  where d is the number of estimated parameters
- Model with lowest AIC will have the lowest error on an independent test set

- Only useful for model selection, not performance estimate
- Only applicable for regression models
- Several variations (BIC, QIC)

Shrinkage and selection

- Ridge regression
  - sum of the squares of coefficients ≤ constant.
- Lasso
  - sum of absolute value of coefficients ≤ constant.
  - Some coefficients are reduced to 0 (effectively leading to variable selection)

Bootstrapping to obtain honest estimates while using **all the data**

1. Draw sample (*with replacement)* of size *N* from data set
2. Build model on this bootstrap sample
3. Compute error on bootstrap sample and on original sample

- The difference between these errors is a (shaky) estimate of the optimism
- The whole procedure is repeated many (e.g. 1,000) times to obtain a stable estimate
- Drawback: computationally intensive

Davison AC, Hinkley DV. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

Note on bootstrapping

- Does not choose complexity
- Should decide on strategy with controlling complexity and apply this also within the bootstrap samples

Types of validity

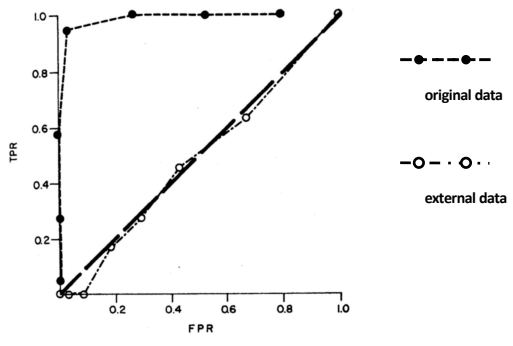
- Internal validity (this is what we looked at)
  The model is valid for patients from the same population and in the same setting
- Prospective validity
  The model is valid for *future* patients from the same population and in the same setting
- External validity
  The model is valid for patients from *another population* or *another setting.*

Justice AC, et al. *Annals of Internal Medicine* 1999; **130**:515–24.

Example: external validity

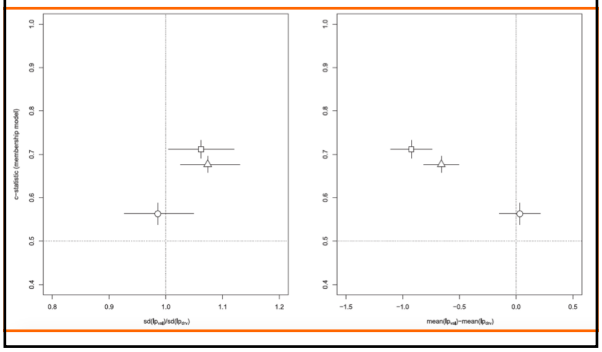
- Prognostic model to predict relapse in patients with acute asthma developed by Fischl et al.
- Based on data from ER patients in Miami
- Useful for "prophylactic" treatment of high-risk patients
- Reported 95% sensitivity and 97% specificity
- Dramatic drop in accuracy when externally validated on patients in Virginia

Centor RM et al. *NEJM* 1984;310(9):577-580.



Interpreting External Validation

Debray, J Clin Epidemiol. 2015



Interpretation

- Reproducibility
- Transportability

Overview

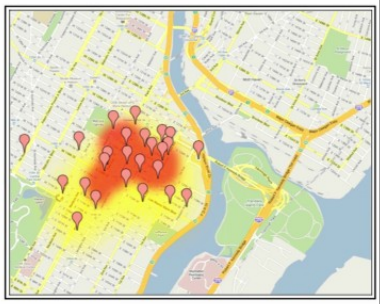
1. Introduction
2. Performance measures
3. Model selection and validation

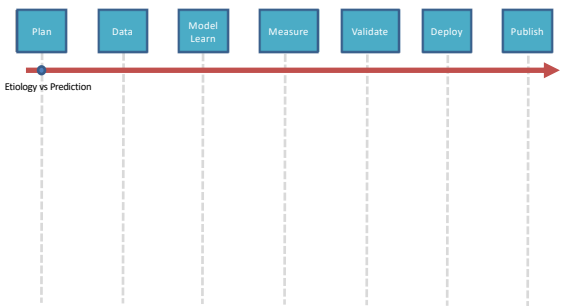
4. Challenges

Opportunities

- Lots of online data
- Various forms of data
  - Demography, clinical, lab, ..
  - Images
  - Sensors
  - Text
- Powerful computing environments
  - Deep learning

Link data: Propeller Health





Best predictions?
understand (causal) relationships?

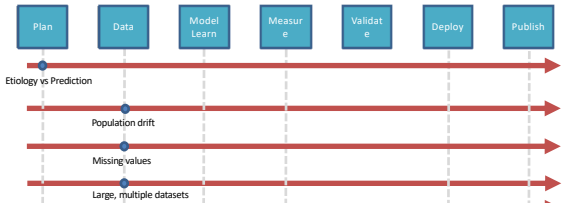
Pitfall: fit for best prediction and interpret as causal model

Matching performance measures to model use

Pitfall: ignoring intended model use when selecting the performance measures

1. informing patients and their families — calibration
2. identification of groups with highest risk — discrimination
3. case-mix correction for quality assessment — calibration

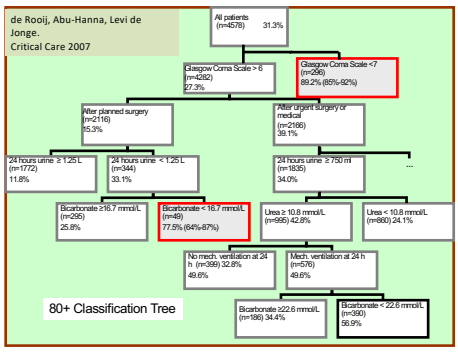
Pitfall only report AUC when calibration is essential.



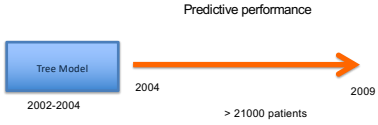
Population drift: Shelf life of prediction models?

- Patients and treatments change over time (Population drift)
- What happens to these results if the predictions change over time?

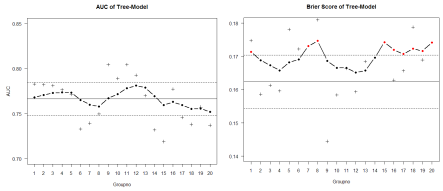
Pitfall: ignoring time and its effects



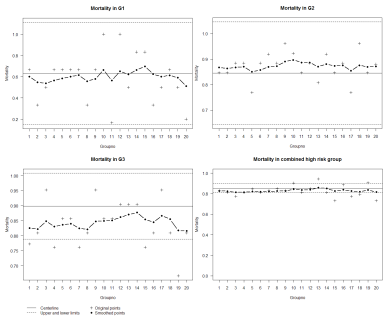
Prospective performance validation



Model performance in SPC



High risk probabilities over time



Implications over time for fitness for use

- For benchmarking the model should not be used
- Useful for identifying patients at high risk

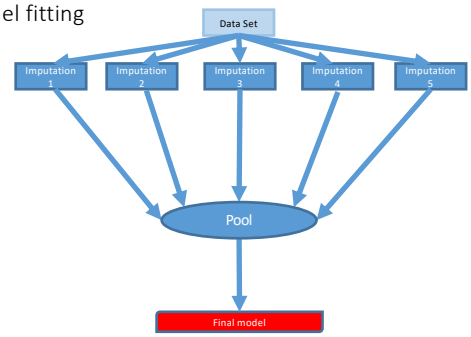
Missing data

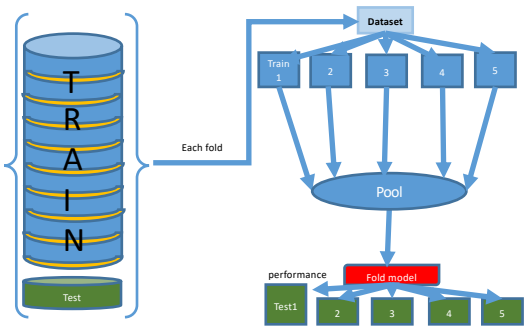
- Completely at random
- At random (predictable from other variables)
- Informative

- How to fit model?
- How to perform internal validation

Pitfall: validating without taking into account decisions made in the model development phase.

Model fitting





Train 1 and Test1 etc are first imputed together and then split

Correlated data

- Many observations obtained from same subject (patient, hospital, …)
- We cannot treat them as if they were iid (independent and identically distributed).
- We could use mixed-effects models (there are new approaches in ML to cope with this).

Pitfall: not taking into account corelated data.

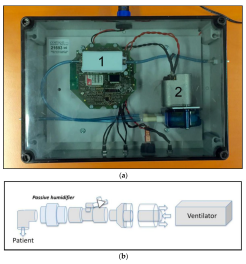
Messy sensor data

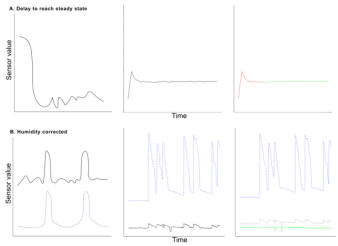
Pitfall: you believe someone telling you they have nice sensor data.

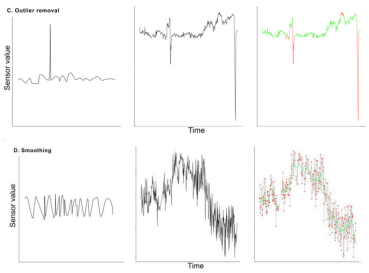
Factors Influencing Continuous Breath Signal in Intubated and Mechanically-Ventilated Intensive Care Unit Patients Measured by an Electronic Nose

Jan Hendrik Leopold [1,2,*], Ameen Abu-Hanna [2], Camilla Colombo [2], Peter J. Sterk [3], Marcus J. Schultz [1] and Lieuwe D. J. Bos [1,3]

Sensors 2016







Big (sensor) data

Pitfall: you believe you can run your data on your laptop.

Clinical and cost-effectiveness of home-based cardiac rehabilitation compared to conventional, centre-based cardiac rehabilitation: Results of the FIT@Home study

Jos J Kraal, M Elske Van den Akker-Van Marle, Ameen Abu-Hanna, Wim Stut, Niels Peek, Hareld MC Kemps

First Published May 23, 2017 | Research Article

Compute cluster

- Ask for resources
- Loop over data files
- Call R module to analyse data

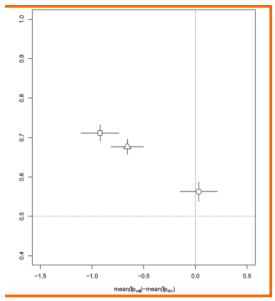
- 12 hours ➔ 3 minutes

Interpreting external validation

Pitfall: you just report external validation without insight.

Interpreting External Validation

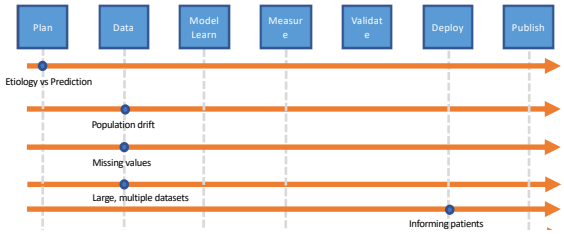
Debray, J Clin Epidemiol. 2015



Interpretation

- Reproducibility
- Transportability

"I want to know why when I'm in a room people ignore me."



Valuable, or merely valid?

Model validation involves more than statistical evaluation of model performance.

There are additional context-dependent criteria that determine a model's use and value

Altman DG, Royston P. *Statistics in Medicine* 2000; **19**:453–73.

Application: futile care in the ICU

Intensivists face the problem of futile care: To treat or not to treat?

Can prognostic models help?

Toma T, Abu-Hanna A, Bosman RJ. J Biomedical Informatics. 2007.
Toma T, Abu-Hanna A, Bosman RJ. Artificial Intelligence in Medicine. 2008.
Toma T, Bosman RJ, Peek N, Siebes A, Abu-Hanna A. J Biomedical Informatics 2010.

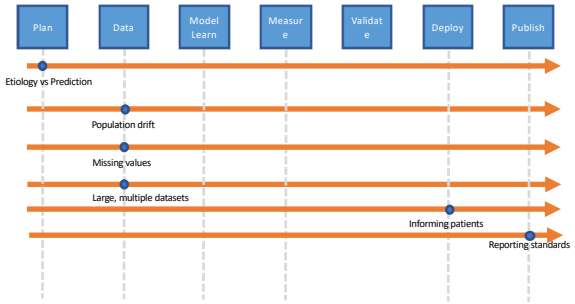
Deployment

Evaluation

- Acceptance by doctors and patients
- Impact on perceptions, confidence, and decisions

You think you're ready to publish

Pitfall: you ignore reporting standards.



Relevant reporting standards

- TRIPOD
  - for development and prediction models
- STROBE and especially RECORD
  - for observational studies and messy data

Conclusion – Prediction models

- Various uses of models with various performance measures
- Methodology from statistics and ML/Computer Science
- Complexity of models is key for obtaining good models
- Many challenges along the way, even with small data!

Stepwise approach for developing prediction models

Step 1: Analysis

- What is the task?
- What is the outcome?
- What is the population we have data from?
- Do you have the relevant predictors?
- Are data reliably collected?
- Missing values?
- What is the EPV (Events per variable)?

Step 2: Coding predictors

- Continuous as linear associations?
- Interaction terms?
- Consider transformations (restricted cubic spline)
- Do not categorize continuous variables (except for getting insight)

Step 3: Model specification and estimation

- Find suitable complexity
  - Stepwise selection of variables? (unstable for low EPV)
  - Use cross validation to decide on complexity
- Include clinical knowledge if available
- Fit model by optimizing some measure, such as likelihood, information gain etc

Step 4: Validity

- Do not trust in-sample performance
- Perform internal validity
  - Cross validation and bootstrapping are better than the split-sample approach
  - Perform temporal/external validation if required
- Provide:
  - Calibration plots
  - Brier score
  - AUC

Step 5: Presentation

- Average predictive comparisons

  For a predictor of interest calculate average:
  - prediction difference at constants v1 and v2
  - prediction difference at actual value and one extra unit
- Formula
- Structure (Tree, NN)
- Nomogram
- Web-based calculator

Step 6: Impact

- Decision-curve analysis
- Effect on perception, confidence, decisions

Methodological and reporting standards

- RECORD
- TRIPOD