Methods for Analytical Validation

William Hsu, PhD
Medical & Imaging Informatics group
Associate Professor of Radiological Sciences, Bioinformatics, and Bioengineering
whsu@mednet.ucla.edu

Learning objectives

- Describe the process of analytical validation
- Explain basic concepts and metrics
- Develop a study for evaluating and comparing algorithms
- Assess the results of an evaluation

Course notes and supplemental materials are available at

https://uclawillhsu.github.io/spie-tutorial/

Learning from varied data types…

Patient ID	Chest Pain	Myocardial Infarct
1234	YES	YES
1235	NO	NO
1236	YES	NO
1237	YES	YES
1238	NO	NO

Textual / numerical

Imaging

Waveform

… to provide actionable insights

Diagnosis?
Clinical deterioration?

Source: https://arxiv.org/pdf/1601.04674.pdf

Source: Ruiwen Ding (UCLA)

Hadjiiski L et al. AAPM task group report 273: Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Med Phys. 2023;50(2):e1–24.

Hadjiiski L et al. AAPM task group report 273: Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Med Phys. 2023;50(2):e1–24.

Driving questions

1. How do we assess whether a model performs as intended?

2. How do we select between multiple models that accomplish the same task?

3. Will the model advance medical decision making?

Source: https://xkcd.com/1838/

Analytical Validation

- Does the learning algorithm correctly process input data to generate accurate, reliable, and precise output data?

Figure 1 - Clinical Evaluation Process

SaMD: Software as a Medical Device

Source: https://www.fda.gov/files/medical%20devices/published/Software-as-a-Medical-Device-%28SAMD%29--Clinical-Evaluation---Guidance-for-Industry-and-Food-and-Drug-Administration-Staff.pdf

Basic concepts: Prediction error

- **Prediction error** can refer to one of two things:
  - In **regression analysis**, it is a measure of how well the model predicts the response variable
    - E.g., root mean squared error
  - In **classification**, it is a measure of how well samples are classified to the correct category
    - E.g., accuracy, precision, recall

Basic concepts: Bias-variance tradeoff

- **Bias** is the difference between the average prediction of our model and the correct value we are trying to predict.
- **Variance** is the model's sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data rather than the intended outputs.

- *High bias & low variance = underfitting*
- *High variance & low bias = overfitting*

Regression example

Evaluate polynomial function that best fits the data

Adapted from https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765

Regression example

Validation results ordered by prediction error

Degrees	MSE
4	0.010549
5	0.010637
7	0.010665
6	0.010887
8	0.011182
3	0.011695
9	0.011757
11	0.011769
10	0.011902
12	0.012642

Adapted from https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765

Training and validation curves

Notes on training and validation curves

Confusion matrix

2-class problems

		Predicted Class	
		Positive	Negative
Real Class	Positive	True Positive (TP)	False Negative (FN) (type 2 error)
	Negative	False Positive (FP) (type I error)	True Negative (TN)

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{true positive rate (recall)} = \frac{TP}{TP + FN}$$

$$\text{false positive rate (1 – specificity)} = \frac{FP}{TN + FP}$$

A single measure doesn't tell the whole story

Precision = 66.7%

Recall = 40%

Precision = 57.1%

Recall = 80%

Accuracy
(both 60%)

	Pos	Neg
Yes	200	100
No	300	400
	P=500	N=500

True	Pos	Neg
Yes	400	300
No	100	200
	P=500	N=500

Precision/recall
(both 66.7% and 40%, respectively)

True class →	Pos	Neg
Yes	200	100
No	300	400
	P=500	N=500

Accuracy = 60%

True class →	Pos	Neg
Yes	200	100
No	300	0
	P=500	N=100

Accuracy = 33%

Source: N Japkowicz. https://www.icmla-conference.org/icmla11/PE_Tutorial.pdf

Receiver operator characteristic (ROC) curve

$$\text{true positive rate (recall)} = \frac{TP}{TP + FN}$$

$$\text{false positive rate (1 – specificity)} = \frac{FP}{TN + FP}$$

ROC curve example

Benign=0; Malignant=1	Radiologist Likert scale (1-5) : 1 (~likely B) 5 (~likely M)	Threshold			
		≥2 (1/0)	≥3 (1/0)	≥4 (1/0)	=5 (1/0)
0	1				
0	2				
0	3				
0	3				
0	4				
1	2				
1	3				
1	4				
1	4				
1	5				

Adapted from Grace H. Kim (UCLA)

Test 1, ≥2	Truth		
	+	–	Total
+ (≥2)			
– (1)			
Total	5	5	10

Sensitivity?
Specificity?

Test	Truth		
	+	–	Total
+ (≥3)			
– (1,2)			
Total	5	5	10

Sensitivity
Specificity

Test	Truth		
	+	–	Total
+ (≥4)			
– (≤3)			
Total	5	5	10

Sensitivity
Specificity

Test	Truth		
	+	–	Total
+ (5)			
– (≤4)			
Total	5	5	10

Sensitivity
Specificity

Benign=0; Malignant=1	Radiologist Likert scale (1-5) : 1 (~likely B) 5 (~likely M)	Threshold			
		≥2 (1/0)	≥3 (1/0)	≥4 (1/0)	=5 (1/0)
0	1	0			
0	2	1			
0	3	1			
0	3	1			
0	4	1			
1	2	1			
1	3	1			
1	4	1			
1	4	1			
1	5	1			

Test 1, ≥2	Truth		
	+	–	Total
+ (≥2)	5	4	9
– (1)	0	1	1
Total	5	5	10

Sensitivity= 100%
Specificity= 20%
1-specificity= 10-20=80%

Test	Truth		
	+	–	Total
+ (≥3)			
– (1,2)			
Total	5	5	10

Sensitivity
Specificity

Test	Truth		
	+	–	Total
+ (≥4)			
– (≤3)			
Total	5	5	10

Sensitivity
Specificity

Test	Truth		
	+	–	Total
+ (5)			
– (≤4)			
Total	5	5	10

Sensitivity
Specificity

Area of trapezoid
= 0.5(L1+L2)*Height

Adapted from Grace H. Kim (UCLA)

Benign=0; Malignant=1	Radiologist Likert scale (1-5) : 1 (~likely B) 5 (~likely M)	Threshold			
		≥2 (1/0)	≥3 (1/0)	≥4 (1/0)	=5 (1/0)
0	1	0	0		
0	2	1	0		
0	3	1	1		
0	3	1	1		
0	4	1	1		
1	2	1	0		
1	3	1	1		
1	4	1	1		
1	4	1	1		
1	5	1	1		

Test 1, ≥2	Truth		
	+	–	Total
+ (≥2)	5	4	9
– (1)	0	1	1
Total	5	5	10

Sensitivity=100%
Specificity=20%
1-specificity=100-20=80%

Test	Truth		
	+	–	Total
+ (≥3)	4	3	7
– (1,2)	1	2	3
Total	5	5	10

Sensitivity= 80%
Specificity =40%
1-specificity=60%

Test	Truth		
	+	–	Total
+ (≥4)			
– (≤3)			
Total	5	5	10

Sensitivity
Specificity

Test	Truth		
	+	–	Total
+ (5)			
– (≤4)			
Total	5	5	10

Sensitivity
Specificity

Area of trapezoid
= 0.5(L1+L2)*Height

Adapted from Grace H. Kim (UCLA)

Benign=0; Malignant=1	Radiologist Likert scale (1-5) : 1 (~likely B) 5 (~likely M)	Threshold			
		≥2 (1/0)	≥3 (1/0)	≥4 (1/0)	=5 (1/0)
0	1	0	0	0	0
0	2	1	0	0	0
0	3	1	1	0	0
0	3	1	1	0	0
0	4	1	1	1	0
1	2	1	0	0	0
1	3	1	1	0	0
1	4	1	1	1	0
1	4	1	1	1	0
1	5	1	1	1	1

Test 1, ≥2	Truth		
	+	–	Total
+ (≥2)	5	4	9
– (1)	0	1	1
Total	5	5	10

Sensitivity= 100%
Specificity= 20%
1-specificity= 100-20=80%

Test	Truth		
	+	–	Total
+ (≥3)	4	3	7
– (1,2)	1	2	3
Total	5	5	10

Sensitivity= 80%
Specificity =40%
1-specificity=60%

Test	Truth		
	+	–	Total
+ (≥4)	3	1	4
– (≤3)	2	4	6
Total	5	5	10

Sensitivity=60%
Specificity=80%
1-specificity: 20%

Test	Truth		
	+	–	Total
+ (5)	1	0	1
– (≤4)	4	5	9
Total	5	5	10

Sensitivity=20%
Specificity=100%
1-specificity=0%

Area Under Curve(AUC) =Sum of trapezoids

AUC= Sum of trapezoids
= 0.5[(0.6+0.2)*0.2
+(0.6+0.8)*0.4
+(0.8+1)*0.2
+(1+1)*.2]=0.74

Adapted from Grace H. Kim (UCLA)

Precision-recall (PR) curves

Estimate	-3.0	-2.5	-1.0	-1.0	-0.5	-0.5	0.5	1.0	2.0	2.0	3.5
Class	-1	-1	-1	-1	-1	+1	-1	+1	-1	+1	+1
Prediction	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Status	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Plots the precision vs. recall as a threshold on the confidence of an instance being positive is varied

Source: https://www.datascienceblog.net/post/machine-learning/interpreting-roc-curves-auc/

Multi-class scenario

Multi-class scenario

		True/Actual		
		Cat	Fish	Hen
Predicted	Cat	4	6	3
	Fish	1	2	0
	Hen	1	2	6

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Precision of 'cat' classifier
= 4/(4+6+3) = 30.8%

Recall of 'cat' classifier
= 4/(4+1+1) = 66.7%

Overall accuracy
= (4+2+6)/25 = 48%

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{true positive rate (recall)} = \frac{TP}{TP + FN}$$

Multi-class scenario

- **No averaging:** Return a measure corresponding to each class.

		True/Actual		
		Cat ( )	Fish ( )	Hen ( )
Predicted	Cat ( )	4	6	3
	Fish ( )	1	2	0
	Hen ( )	1	2	6

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

- Precision of ‘cat’ classifier = 4/(4+6+3) = 30.8%
- Precision of ‘fish’ classifier = 2/(1+2) = 66.7%
- Precision of ‘hen’ classifier = 6/(1+2+6) = 66.7%

Multi-class scenario

- **Macro averaging:** Calculate the measure independently for each class then then find their unweighted mean.

		True/Actual		
		Cat ( )	Fish ( )	Hen ( )
Predicted	Cat ( )	4	6	3
	Fish ( )	1	2	0
	Hen ( )	1	2	6

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

- Macro-precision of classifier is (31% + 67% + 67%) / 3 = 54.7%

Multi-class scenario

- **Micro averaging:** Aggregate contributions of all classes to compute the average measure by counting the total number of times each class was correctly predicted and incorrectly predicted.

		True/Actual		
		Cat ( )	Fish ( )	Hen ( )
Predicted	Cat ( )	4	6	3
	Fish ( )	1	2	0
	Hen ( )	1	2	6

https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

- TP=4+2+6=12
- FP=6+3+1+0+1+2=13
- Micro-precision of classifier is 12/(12+13) = 48%

Other measures of interest

- Comparing segmented regions
  - Overlap measures
    - Dice coefficient
    - Intersection over union (Jaccard)
  - Surface distance
    - Hausdorff distance

$$\mathrm{Dice}(P,T) = \frac{|P_1 \wedge T_1|}{(|P_1| + |T_1|)/2}$$

- Evaluating clustering algorithms
  - Purity
  - Rand index
  - http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

Calibration

- How well does the model's predicted probability approximates the actual event probability

Notes on measures

- Use ROC when both classes detection is equally important
- Use ROC when the positives are the majority
- PRC may be more informative than ROC for clinical decisions → positive predictive value vs false positive rate
- Area under the ROC (AUROC) and area under the PRC (AUPRC) are related -- AUROC may be preferred over AUPRC
  - McDermott et al, 2024: https://arxiv.org/abs/2401.06091
- Micro-average is preferable over macro-average in situations where a class imbalance exists

Evaluation designs

Reference: Japkowicz and Shah. https://doi.org/10.1017/CBO9780511921803.004

Train-validation-test

- Split data into three groups
- Use training data to fit different models
- Use validation data to estimate generalization error (and tune model)
- Use test data to assess the performance

Notes on train-validation-test

- Need to find a balance between the number of training and testing cases
  - A larger training set provides the learning algorithm with a more representative sample of the data
  - A larger test set provides a more reliable estimate of accuracy
- A single training set does not reveal how sensitive the accuracy of the learning algorithm is to a particular training sample → sampling bias
  - Use re-sampling methods to train-validate
- Used to train/test a final model – not a great way to determining model training strategy

Cross validation

- Partition data into k-subsamples then iteratively assign one subsample as the test set, train the learning algorithm on the rest

Nested cross validation

Source: https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7

Stratified cross-validation

When randomly selecting training or validation sets, stratified cross-validation is used to ensure that class proportions are maintained in each selected set

Bootstrapping

- Bootstrapping assumes that available samples are representative and creates a large number of new samples by drawing with replacement from the available samples
- Bootstrapping can be useful when the sample size is too small for cross validation or leave-one-out approaches to yield a good estimate

Repeated k-fold cross-validation

- To obtain more stable estimates of an algorithm's performance, perform multiple runs of simple re-sampling schemes
  - Dietterich (1998) proposed the 5x2CV, in which a two-fold cross-validation is repeated 5 times

Source: https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7

Reference: https://sci2s.ugr.es/keel/pdf/algorithm/articulo/dietterich1998.pdf

Notes about evaluation design

- The goal of cross-validation studies is to identify the optimal modeling approach, not an individual learned model
  - Use cross validation to determine optimal modeling approach
  - Train model using optimal approach on the entire development set
  - Evaluate final model on the held-out test set
- 10-fold cross-validation is commonly used, but smaller values of k can be used when learning takes a lot of time
- When working with time series data, training data must always be selected before the test data (otherwise training data will contain information from the future)

Comparing models: Motivation

- Which model is better?

Algo	Acc	RMSE	TPR	FPR	Prec	Rec	F	AUC	Info S
NB	71.7	.4534	**.44**	.16	.53	**.44**	**.48**	**.7**	48.11
C4.5	**75.5**	**.4324**	.27	**.04**	**.74**	.27	.4	.59	34.28
3NN	72.4	.5101	.32	.1	.56	.32	.41	.63	43.37
Ripp	71	.4494	.37	.14	.52	.37	.43	.6	22.34
SVM	69.6	.5515	.33	.15	.48	.33	.39	.59	**54.89**
Bagg	67.8	.4518	.17	.1	.4	.17	.23	.63	11.30
Boost	70.3	.4329	.42	.18	.5	.42	.46	**.7**	34.48
RanF	69.23	.47	.33	.15	.48	.33	.39	.63	20.78

Reference: Japkowicz and Shah. https://doi.org/10.1017/CBO9780511921803.004

What is being tested?

Are the performance measurements attributable to the machine learning algorithm or are they observed by chance?

- Provide evidence that our selected evaluation measure is representative of the general behavior of our machine learning algorithm

- Parametric versus nonparametric test
  - Parametric tests make strong assumptions about the distribution of the underlying data (normal, Bernoulli, etc) but require smaller sample size
  - Nonparametric tests make weaker assumptions (skewed distribution, presence of outliers) but are also less powerful (less apt at rejecting the null hypothesis when it is false)

Considerations

- Can a parametric versus nonparametric test be used
- Select appropriate test based on the following scenarios
  - The comparison of 2 algorithms on the same dataset
  - The comparison of 2 algorithms on different datasets
  - The comparison of multiple algorithms on multiple datasets

Test selection

Reference: Japkowicz and Shah. https://doi.org/10.1017/CBO9780511921803.004

The paired t-test

- Given two matched samples
  - the results of two classifiers applied to the same data set with **matching randomizations** and **partitions**
- Test whether the difference in **mean of a metric (e.g., accuracy)** between these two classifiers is significant
  - Test whether the two samples come from the same population.
  - Look at the distribution in **observed means** and **standard deviation**.
- We assume that the difference between means is zero (the null hypothesis) and see if we can reject this hypothesis.

Notes on using the t-test

- **Normality:** The t-test requires that the samples come from normally distributed population.
- **Randomness:** The sample should be representative of the underlying population.
- **Equal variance**: The two samples come from populations with equal variance.
- Measures effect, but not the size of the effect → Pearson correlation coefficient

	Classifier A	Classifier B	Difference
Iteration 1	P_{A1}	P_{B1}	$d_1 = P_{A1} - P_{B1}$
Iteration 2	P_{A2}	P_{B2}	$d_2 = P_{A2} - P_{B2}$
Iteration 3	P_{A3}	P_{B3}	$d_3 = P_{A3} - P_{B3}$
Iteration 4	P_{A4}	P_{B4}	$d_4 = P_{A4} - P_{B4}$
	⋮	⋮	⋮
Iteration n-1	$P_{A(n-1)}$	$P_{B(n-1)}$	$d_{n-1} = P_{A(n-1)} - P_{B(n-1)}$
Iteration n	P_{An}	P_{Bn}	$d_n = P_{Bn} - P_{Bn}$

P_{Ai}: Performance of classifier A at iteration "i"

P_{Bi}: Performance of classifier B at iteration "i"

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$$

Reference: Japkowicz and Shah. https://doi.org/10.1017/CBO9780511921803.004

McNemar's test

- Non-parametric counterpart of the t-test observed on the test set
  - # of instances misclassified by both classifiers C_{00}
  - # of instances misclassified by first classifier but correctly classified by the second classifier C_{01}
  - # of instances misclassified by the second classifier but correctly classified by the first classifier C_{10}
  - # of instances correctly classified by both classifiers C_{11}

		Classifier f_2	
		0	1
Classifier f_1	0	c_{00}^{Mc}	c_{01}^{Mc}
	1	c_{10}^{Mc}	c_{11}^{Mc}

$$X_{MC}^2 = \frac{(|C_{01} - C_{10}| - 1)^2}{C_{01} + C_{10}}$$

Notes on using the McNemar's test

- Similar to a t-test on dichotomous rather than continuous values
- Does not require the same assumptions about the distribution as the t-test
- Ideally applied over an independent test set for performance assessment
- The two classifiers should use the same training and test sets during each fold of the cross-validation run
- C_{01} + C_{10} cannot be too "small" → does not follow X^2 distribution and must use a binomial distribution instead (sign test)

The Sign test

- The sign test can be used either to compare
  - two classifiers on a single domain (using the results at each fold as a trial)
  - two classifiers on multiple domains
- We count
  - number of times that algorithm 1 outperforms algorithm 2, n_{a1}
  - number of times that algorithm 2 outperforms algorithm 1, n_{a2}
- The null hypothesis (stating that the two classifiers perform equally well) holds if the number of wins follows a binomial distribution.
- Practically speaking, a classifier should perform better on at least w_α datasets to be considered statistically significantly better at the α significance level, where w_α is the critical value for the sign test at the α significance level

Reference: Japkowicz and Shah. https://doi.org/10.1017/CBO9780511921803.004≈

Notes on using the sign test

Dataset	NB	SVM	Adaboost	Rand Forest
Anneal	96.43	**99.44**	83.63	**99.55**
Audiology	73.42	**81.34**	46.46	**79.15**
Balance Scale	72.30	**91.51**	72.31	**80.97**
Breast Cancer	**71.70**	66.16	**70.28**	69.99
Contact Lenses	**71.67**	**71.67**	**71.67**	**71.67**
Pima Diabetes	74.36	**77.08**	74.35	**74.88**
Glass	**70.63**	62.21	44.91	**79.87**
Hepatitis	**83.21**	80.63	82.54	**84.58**
Hypothyroid	**98.22**	93.58	93.21	**99.39**
Tic-Tac-Toe	69.62	**99.90**	72.54	**93.94**

- NB vs SVM: n_{f1}= 4.5, n_{f2}= 5.5 and $w_{0.05}$ = 8 ➔ we cannot reject the null hypothesis stating that NB and SVM perform similarly on these data sets for level α=0.05 (1-tailed)
- Ada vs RF: n_{f1}=1.5 and n_{f2}=8.5 ➔ we can reject the null hypothesis at level α=0.05 (1-tailed) and conclude that RF is significantly better than Ada on these data sets at that significance level.

Notes on using the sign test

A.4 The Table of Critical Values for the Signed Test

Table A.4. *Critical values of T for the sign test*

Level of significance α					Level of significance α				
Two-sided	0.10	0.05	0.02	0.01	Two-sided	0.10	0.05	0.02	0.01
One-sided	0.05	0.025	0.01	0.005	One-sided	0.05	0.025	0.01	0.005
n					n				
1	–	–	–	–	31	11	13	15	17
2	–	–	–	–	32	12	14	16	16
3	–	–	–	–	33	11	13	15	17
4	–	–	–	–	34	12	14	16	16
5	5	–	–	–	35	11	13	15	17
6	6	6	–	–	36	12	14	16	18
7	7	7	7	–	37	11	13	17	17
8	6	8	8	8	38	12	14	16	18
9	7	7	9	9	39	13	15	17	17
10	8	8	10	10	40	12	14	16	18
11	7	9	9	11	45	13	15	17	19
12	8	8	10	10	46	14	16	18	20
13	7	9	11	11	49	13	15	19	19
14	8	10	10	12	50	14	16	18	20

Reference: Japkowicz and Shah. https://doi.org/10.1017/CBO9780511921803.004≈

Wilcoxon's Signed-Rank Test

- Non-parametric
- For each domain, we calculate the difference in the performance of the two classifiers.
  - Rank the absolute values of these differences and graft the signs in front of the ranks.
  - Calculate the sum of positive and negative ranks, respectively (W_{S1} and W_{S2})
  - $T_{Wilcox} = \min(W_{S1}, W_{S2})$
- Compare to critical value V_α. If $V_\alpha \geq T_{Wilcox}$ we reject the null hypothesis that the performance of the two classifiers is the same, at the α confidence level.

Reference: Japkowicz and Shah. https://doi.org/10.1017/CBO9780511921803.004

Wilcoxon's signed-Rank Test: Illustration

Data	NB	SVM	NB-SVM	|NB-SVM|	Ranks	± Ranks
1	.9643	.9944	-0.0301	0.0301	3	-3
2	.7342	.8134	-0.0792	0.0792	6	-6
3	.7230	.9151	-0.1921	0.1921	8	-8
4	.7170	.6616	+0.0554	0.0554	5	+5
5	.7167	.7167	0	0	Remove	Remove
6	.7436	.7708	-0.0272	0.0272	2	-2
7	.7063	.6221	+0.0842	0.0842	7	+7
8	.8321	.8063	+0.0258	0.0258	1	+1
9	.9822	.9358	+0.0464	0.0464	4	+4
10	.6962	.9990	-0.3028	0.3028	9	-9

W_{S1} = 17 and W_{S2} = 28 ➔ T_{Wilcox} = min(17, 28) = 17
For n= 10-1 degrees of freedom and α = 0.005, V = 8 for the 1-sided test. V must be larger than T_{Wilcox} in order to reject the hypothesis. Since 17 > 8, we cannot reject the hypothesis that NB's performance is equal to that of SVM at the 0.005 level.

Analysis of variance (ANOVA)

- For the case of multiple classifiers and multiple domains, two alternatives are possible
  - The parametric alternative is (one-way repeated measure) ANOVA
  - The non-parametric alternative is Friedman's Test
- H_0: all the classifiers perform equally
- H_A: at least one pair of classifiers with significantly different performance
- If H_0 is rejected, the test is followed by a post-hoc test whose job is to identify the significantly different pairs of classifiers

Notes about statistical testing

- Keep in mind the assumptions set by the selected statistical test (e.g., underlying distribution)
- Failure to achieve a statistically significant result <u>should not</u> be interpreted as a true lack of difference – especially when the study is statistically underpowered
  - Statistical significance <u>does not</u> imply that the result is clinically meaningful
- Multiple hypothesis testing should be accounted for
  - When multiple statistical hypotheses are tested using the same dataset, the chance of observing a rare event increase, increasing the likelihood of incorrectly concluding that a real effect has been observed
  - E.g., Bonferroni correction

Guidelines for reporting data

- Data sources
  - Inclusion criteria
  - Selection of data subsets
- Data pre-processing steps
  - Definition of data elements
  - De-identification methods
  - How missing data were handled
- Reference standard
  - How was the standard defined?
  - The rationale for choosing the standard
  - Source of ground truth annotations
    - If generated by human annotators, were inter and intra-rater variability mitigated?

Data sources

Park S, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology. 2018;286(3):800–9.

Reference standard

- Source of the standard
  - Completeness?
  - Consistent?
- Rationale for choosing the standard
- How was the standard defined?
  - Subject to measurement error?
  - Subjective judgment?

	Majority Decision of Retinal Specialist Grading before Adjudication				
	No	*Mild*	*Moderate*	*Severe*	*Proliferative*
Adjudicated Consensus					
No	1469	4	5	0	0
Mild	58	62	5	0	0
Moderate	22	3	118	1	0
Severe	0	0	13	36	1
Proliferative	0	0	0	1	15

Confusion matrix for diabetic retinopathy between the grade determined by majority decision and adjudicated consensus.

Source: https://doi.org/10.1016/j.ophtha.2018.01.034

Guidelines for reporting models

- Model
  - Description of model
  - Inputs and outputs
  - Hyperparameters
- Implementation
  - Software libraries and frameworks
  - Computational requirements
- Training
  - How was the dataset split for training?
  - Was transfer learning used?
  - Method of selecting final model

Guidelines for reporting evaluation

- Evaluation approach
  - Metrics used to measure model performance
  - Statistical tests used
  - Whether sensitivity analysis was performed to evaluate model robustness
  - Techniques for explaining model results, if applicable
  - Validation of model using external data
- Results
  - Whether performance metrics were compared with previously reported models, if possible
  - Whether confidence intervals are provided
  - Failure analysis
- Limitations
  - Possible sources of bias
  - Implications for practice

Helpful reporting standards

- Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)
  - https://www.equator-network.org/reporting-guidelines/tripod-statement/
  - TRIPOD-AI under development: https://bmjopen.bmj.com/content/11/7/e048008 | https://osf.io/zyacb/
- Prediction model Risk Of Bias ASsessment Tool (PROBAST)
  - https://www.acpjournals.org/doi/10.7326/M18-1376
  - PROBAST-AI under development: https://bmjopen.bmj.com/content/11/7/e048008
- Reporting of clinical trials involving AI
  - CONSORT/CONSORT-AI: https://www.nature.com/articles/s41591-020-1034-x
  - SPIRIT (protocol reporting): https://www.nature.com/articles/s41591-020-1037-7

Notes on external validation

- New breast AI/ML algorithms are becoming commercially available and are being reported to have performance on par with breast radiologists
  - Lack of transparency about the cohorts in which the algorithm was trained and evaluated
    → will the performance generalize?
  - Guidelines on how institutions should review and adopt AI/ML algorithms are sparse
    → what factors should institutions consider before buying?
  - Check if the AI/ML algorithm works as intended
    → when does an algorithm deviate from stated performance?
- Objective: Perform an external validation of the ”DREAM Challenge Competition Ensemble Method” model on a retrospective cohort from our institution
  - Primary measures: AUC, sensitivity, specificity
  - Identify conditions when model performance differs from previously reported results

Source: http://jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2022.42343

Notes on external validation

Notes on external validation

- Source of reference: outcomes from the regional cancer registry

Source: http://jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2022.42343

Notes on external validation

Characteristic	Full cohort, unweighted	Analyzed subset[a] Unweighted, No. (%)
Women		
Total, No.	41 343	26 817
Diagnosed with BC within 12 mo of mammogram	714 (1.7)	573 (2.1)
Screening examinations		
Total, No.	121 753	37 317
BI-RADS category		
0	8493 (6.9)	3809 (10.2)
1	46 093 (37.9)	14 063 (37.7)
2	66 632 (54.7)	19 315 (51.8)
3, 4, or 5	535 (0.4)	130 (0.3)
BC diagnosed within 12 mo	723 (0.6)	576 (1.5)
Invasive BC	576 (79.7)	462 (80.2)
DCIS	147 (20.3)	114 (19.8)
No BC diagnosed within 12 mo	121 030 (99.4)	36 741 (98.5)
Age, mean (SD), y	58.9 (11.6)	58.4 (11.5)
BMI, mean (SD)	28.4 (6.9)	29.0 (7.0)
Personal history of breast cancer[b]	3751 (6.1)	1864 (6.1)
Race and ethnicity[c]		
Asian	10 913 (10.3)	3338 (9.7)
Black	9558 (9.0)	2972 (8.6)
Hispanic	11 449 (10.8)	3699 (10.6)
White	61 703 (58.3)	20 602 (59.3)
Other[d]	12 254 (11.6)	4093 (11.8)

Source: http://jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2022.42343

Notes on external validation

Source: http://jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2022.42343

Resources

- Electronic course notes and supplementary materials
  - https://uclawillhsu.github.io/spie-tutorial/
- Evaluating learning algorithms (Japkowicz and Shah)
  - https://doi.org/10.1017/CBO9780511921803
- Chapter 10: Evaluation (Medical Imaging Informatics, Bui and Taira, eds. 2010)
  - https://link.springer.com/chapter/10.1007/978-1-4419-0385-3_10

Acknowledgements and references

William Hsu, PhD
Professor of Radiological Sciences, Bioinformatics, and Bioengineering
whsu@mednet.ucla.edu
http://hsu-lab.com

Research team (Hsu Lab @ Medical & Imaging Informatics)

The Hsu Lab gratefully acknowledges support from the National Science Foundation (#1722516), National Institutes of Health (R01 EB031993; R01 CA271034; R01 HL127153; R01 CA22079; R01 EB029346; R01 CA210360; R37 CA240403), UCLA SPORE in Prostate Cancer, Jonsson Comprehensive Cancer Center, and the Department of Radiological Sciences.

Brief read

- Liu Y, Chen PH, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. Jama. 2019 Nov 12;322(18):1806-16.

More in-depth

- Steyerberg, EW. Evaluation of Performance. In: Clinical Prediction Models. Statistics for Biology and Health. Springer, Cham; 2019.

Textbook reference

- Japkowicz N, Shah M. Evaluating learning algorithms: a classification perspective. Cambridge University Press; 2011.