Digital Transformation of Healthcare

Evaluating Predictions & Data Quality

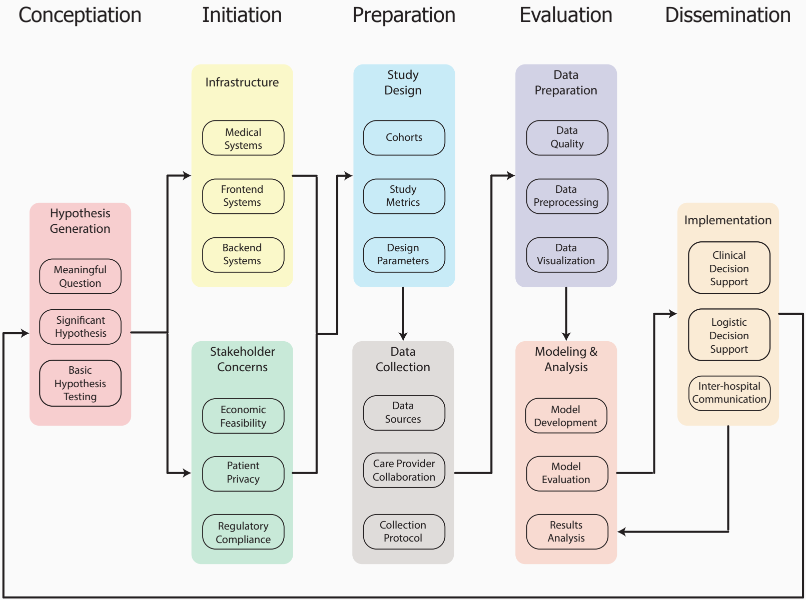
Michoel Snow, MD PhD, Glen Ferguson, PhD

Center for Health Data Innovations

Objectives

After this lecture students will be able to

- Calculate common classification and regression metrics
- Describe the role of simple classification metrics
- Evaluate the implementation of metrics for a study
- Articulate the information underlying common compound classification metrics
- Classify regression metrics
- Connect regression metric outcomes to facets of the associated models
- Identify transition points which can affect data quality
- Discuss methods for measuring and evaluating data quality



Conception
Initiation
Preparation
Evaluation
Dissemination
Hypothesis Generation
Meaningful Question
Significant Hypothesis
Basic Hypothesis Testing
Infrastructure
Medical Systems
Frontend Systems
Backend Systems
Stakeholder Concerns
Economic Feasibility
Patient Privacy
Regulatory Compliance
Study Design
Cohorts
Study Metrics
Design Parameters
Data Collection
Data Sources
Care Provider Collaboration
Collection Protocol
Data Preparation
Data Quality
Data Preprocessing
Data Visualization
Modeling & Analysis
Model Development
Model Evaluation
Results Analysis
Implementation
Clinical Decision Support
Logistic Decision Support
Inter-hospital Communication

Metrics for Evaluation of Classification Models

Clinical Cases

Case	TP	FP	TN	FN
LDCT[1]	649	17,497	49,792	5,532
AAA [2]	600	734	25,480	61
HTN[3]	17	6	65	14

[1] National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine, 365(5), 395-409.

[2] Thompson, S. G., Ashton, H. A., Gao, L., Buxton, M. J., Scott, R. A. P., & Multicentre Aneurysm Screening Study (MASS) Group. (2012). Final followup of the Multicentre Aneurysm Screening Study (MASS) randomized trial of abdominal aortic aneurysm screening. British Journal of Surgery, 99(12), 1649-1656.

[3] Stergiou, G. S., Nasothimiou, E., Giovas, P., Kapoyiannis, A., & Vazeou, A. (2008). Diagnosis of hypertension in children and adolescents based on home versus ambulatory blood pressure monitoring. Journal of hypertension, 26(8), 1556-1562.

Clinical Cases

Case	TP	FP	TN	FN
LDCT[1]	649	17,497	49,792	5,532
AAA [2]	600	734	25,480	61
HTN[3]	17	6	65	14

- Are these *good* tests?
- In what contexts are they useful?
- For which metrics are they misleading?

[1] National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine, 365(5), 395-409.

[2] Thompson, S. G., Ashton, H. A., Gao, L., Buxton, M. J., Scott, R. A. P., & Multicentre Aneurysm Screening Study (MASS) Group. (2012). Final followup of the Multicentre Aneurysm Screening Study (MASS) randomized trial of abdominal aortic aneurysm screening. British Journal of Surgery, 99(12), 1649-1656.

[3] Stergiou, G. S., Nasothimiou, E., Giovas, P., Kapoyiannis, A., & Vazeou, A. (2008). Diagnosis of hypertension in children and adolescents based on home versus ambulatory blood pressure monitoring. Journal of hypertension, 26(8), 1556-1562.

Confusion Matrix

actual outcome

		p	**n**	
predicted outcome	**p′**	TP	FP	$\text{PPV} = \dfrac{TP}{TP + FP}$
	n′	FN	TN	$\text{NPV} = \dfrac{TN}{FN + TN}$
		$\text{Sens} = \dfrac{TP}{TP + FN}$	$\text{Spec} = \dfrac{TN}{FP + TN}$	$\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN}$

Confusion Matrix

		actual outcome p	actual outcome n	
predicted outcome	p′	TP	FP	$PPV = \frac{TP}{TP+FP}$
	n′	FN	TN	$NPV = \frac{TN}{FN+TN}$
		$Sens = \frac{TP}{TP+FN}$	$Spec = \frac{TN}{FP+TN}$	$Acc = \frac{TP+TN}{TP+TN+FP+FN}$

Case	TP	FP	TN	FN	Sens	Spec	PPV	NPV	Acc
LDCT	649	17,497	49,792	5,532					
AAA	600	734	25,480	61					
HTN	17	6	65	14					

Estimate if you think the value will be low, medium or high

Confusion Matrix

		actual outcome		
		p	**n**	
predicted outcome	$\mathbf{p'}$	TP	FP	$\text{PPV} = \frac{TP}{TP + FP}$
	$\mathbf{n'}$	FN	TN	$\text{NPV} = \frac{TN}{FN + TN}$
		$\text{Sens} = \frac{TP}{TP + FN}$	$\text{Spec} = \frac{TN}{FP + TN}$	$\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN}$

Case	TP	FP	TN	FN	Sens	Spec	PPV	NPV	Acc
LDCT	649	17,497	49,792	5,532	10	74	4	90	69
AAA	600	734	25,480	61	91	97	45	100	97
HTN	17	6	65	14	55	92	74	82	80

Confusion Matrix

		actual outcome		
		$\mathbf{p}$	$\mathbf{n}$	
predicted outcome	$\mathbf{p}'$	TP	FP	$\mathrm{PPV} = p\left(\mathbf{p} \mid \mathbf{p}'\right)$
	$\mathbf{n}'$	FN	TN	$\mathrm{NPV} = p\left(\mathbf{n} \mid \mathbf{n}'\right)$
		$\mathrm{Sens} = p\left(\mathbf{p}' \mid \mathbf{p}\right)$	$\mathrm{Spec} = p\left(\mathbf{n}' \mid \mathbf{n}\right)$	$\mathrm{Acc} = p\left(TP + TN\right)$

Confusion Matrix

		actual outcome **p**	**n**	
predicted outcome	$\mathbf{p}'$	TP	FP	$\text{PPV} = p(\mathbf{p} \mid \mathbf{p}')$
	$\mathbf{n}'$	FN	TN	$\text{NPV} = p(\mathbf{n} \mid \mathbf{n}')$
		$\text{Sens} = p(\mathbf{p}' \mid \mathbf{p})$	$\text{Spec} = p(\mathbf{n}' \mid \mathbf{n})$	$\text{Acc} = p(TP + TN)$

Case	**TP**	**FP**	**TN**	**FN**	**Sens**	**Spec**	**PPV**	**NPV**	**Acc**
LDCT	649	17,497	49,792	5,532	10	74	4	90	69
AAA	600	734	25,480	61	91	97	45	100	97
HTN	17	6	65	14	55	92	74	82	80

Parameter	Interpretation	Appropriate for
Accuracy	Overall proximity of test to reality	Balanced sample sizes
Sensitivity		
Specificity		
PPV		
NPV		

Confusion Matrix

		actual outcome **p**	**n**	
predicted outcome	$\mathbf{p}'$	TP	FP	PPV $= p\,(\mathbf{p} \mid \mathbf{p}')$
	$\mathbf{n}'$	FN	TN	NPV $= p\,(\mathbf{n} \mid \mathbf{n}')$
		Sens $= p\,(\mathbf{p}' \mid \mathbf{p})$	Spec $= p\,(\mathbf{n}' \mid \mathbf{n})$	Acc $= p\,(TP + TN)$

Case	**TP**	**FP**	**TN**	**FN**	**Sens**	**Spec**	**PPV**	**NPV**	**Acc**
LDCT	649	17,497	49,792	5,532	10	74	4	90	69
AAA	600	734	25,480	61	91	97	45	100	97
HTN	17	6	65	14	55	92	74	82	80

Parameter	Interpretation	Appropriate for
Accuracy	Overall proximity of test to reality	Balanced sample sizes
Sensitivity	Chance of a false negative	Cheap testing/Severe disease
Specificity	Chance of a false positive	Expensive testing/Mild disease
PPV	Sensitivity diagnostic utility	Balanced prevalence
NPV	Specificity diagnostic utility	Balanced prevalence

Combined Statistics

Function of	Metric	Formula
Sensitivity, Specificity	Positive Likelihood Ratio/ROC	$\frac{sensitivity}{1 - specificity}$
Sensitivity, Specificity	Negative Likelihood Ratio	$\frac{1 - sensitivity}{specificity}$
Sensitivity, PPV	F1 score	$\frac{2}{\frac{1}{sensitivity} + \frac{1}{PPV}}$
TP, TN, FP, FN	Matthews correlation coefficient	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

Combined Statistics

Function of	Metric	Formula
Sensitivity, Specificity	**Positive Likelihood Ratio/ROC**	$\frac{sensitivity}{1 - specificity}$
Sensitivity, Specificity	**Negative Likelihood Ratio**	$\frac{1 - sensitivity}{specificity}$
Sensitivity, PPV	**F1 score**	$\frac{2}{\frac{1}{sensitivity} + \frac{1}{PPV}}$
TP, TN, FP, FN	Matthews correlation coefficient	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

Likelihood Ratios

$$LR+ = \frac{\mathit{sensitivity}}{1 - \mathit{specificity}} = \frac{P\left(T+ \mid D+\right)}{P\left(T+ \mid D-\right)}$$

$$LR- = \frac{1 - \mathit{sensitivity}}{\mathit{specificity}} = \frac{P\left(T- \mid D+\right)}{P\left(T- \mid D-\right)}$$

Likelihood Ratios

$$LR+ = \frac{sensitivity}{1 - specificity} = \frac{P\left(T+ \mid D+\right)}{P\left(T+ \mid D-\right)}$$

$$LR- = \frac{1 - sensitivity}{specificity} = \frac{P\left(T- \mid D+\right)}{P\left(T- \mid D-\right)}$$

Does a test result change the probability that a person has a certain condition?

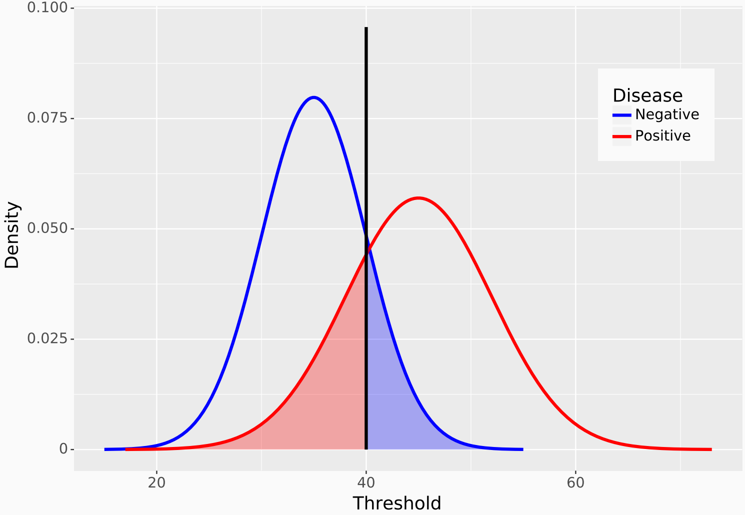
Likelihood Ratios

Likelihood Ratio	Approximate Change in Probability(%)
0.1	-45
0.2	-30
0.5	-15
1	0
2	+15
5	+30
10	+45

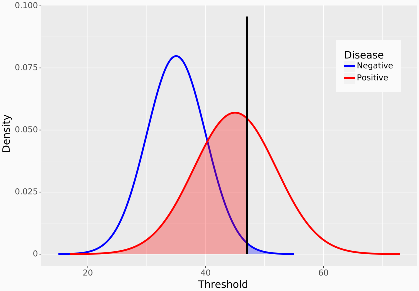
Change in post test probability $\approx 0.2 \times \ln LR$ [4]

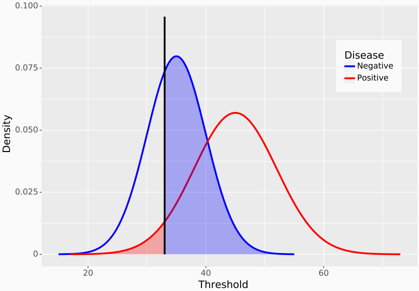
[4] McGee, Steven. "Simplifying likelihood ratios." Journal of general internal medicine 17.8 (2002): 647-650. APA

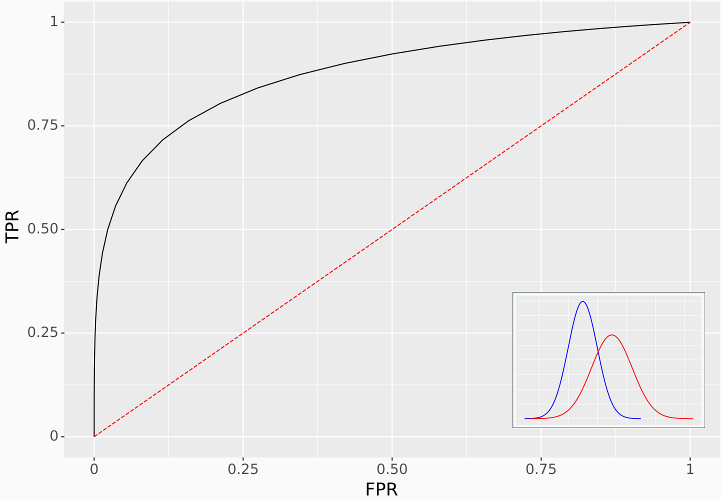
Hypothesis Testing



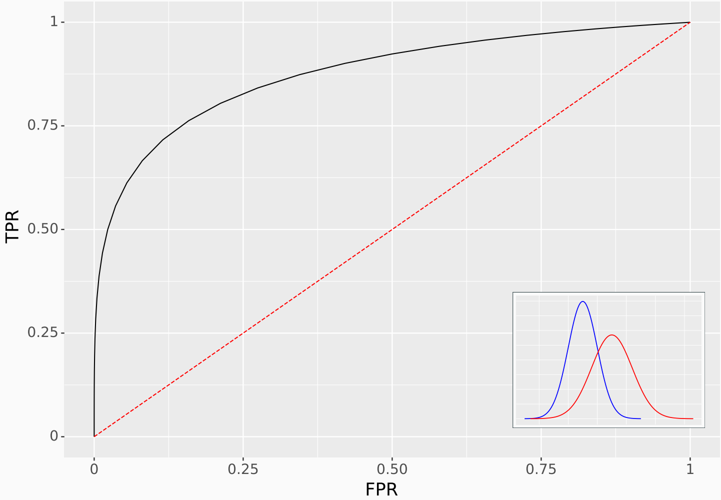
Discrimination Thresholds






1
0.75
0.50
0.25
0
TPR
0
0.25
0.50
0.75
1
FPR

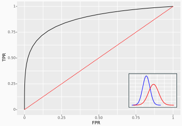
Receiver Operating Characteristic



What does the Area Under the Curve (AUC) correspond to?

What does the diagonal correspond to?

Receiver Operating Characteristic



What does the Area Under the Curve (AUC) correspond to?
What does the diagonal correspond to?

1. Given a positive test result what are the chances that the subject is truly positive irrespective of prevalence?
2. PPV is threshold dependent, while AUC is threshold independent but variable dependent

F1 Score

			actual outcome **p**	**n**	
predicted outcome		$\mathbf{p}'$	TP	FP	PPV $= p\,(\mathbf{p} \mid \mathbf{p}')$
		$\mathbf{n}'$	FN	TN	NPV $= p\,(\mathbf{n} \mid \mathbf{n}')$
			Sens $= p\,(\mathbf{p}' \mid \mathbf{p})$	Spec $= p\,(\mathbf{n}' \mid \mathbf{n})$	Acc $= p\,(TP + TN)$

$$F1 = \frac{2}{\frac{1}{sensitivity} + \frac{1}{PPV}} = 2 \times \frac{PPV \cdot sensitivity}{PPV + sensitivity}$$

How does F1 differ from AUC?

F1 Score

predicted outcome	actual outcome: p	actual outcome: n	
p′	TP	FP	PPV $= p(\mathbf{p} \mid \mathbf{p}')$
n′	FN	TN	NPV $= p(\mathbf{n} \mid \mathbf{n}')$
	Sens $= p(\mathbf{p}' \mid \mathbf{p})$	Spec $= p(\mathbf{n}' \mid \mathbf{n})$	Acc $= p(TP + TN)$

$$F1 = \frac{2}{\frac{1}{sensitivity} + \frac{1}{PPV}} = 2 \times \frac{PPV \cdot sensitivity}{PPV + sensitivity}$$

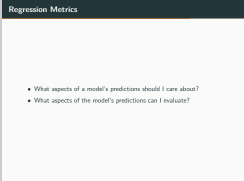
How does F1 differ from AUC?

1. F1 is sensitivity modified by prevalence. So a low prevalence will hurt your F1 score but might not affect your AUC. F1 is threshold specific and corresponds to a point on the ROC curve
2.

Metrics for Evaluation of Regression Models

Regression Metrics

- What aspects of a model's predictions should I care about?
- What aspects of the model's predictions can I evaluate?



1. – Accuracy (bias), precision (variance)
   – Average distance of errors
   – Worst case error
   – Do large errors matter more than small errors
   – Maximal distance of errors
   – Difference between my model and some standard model
2. difference, squared difference, min/max, variance of predictions, relative difference (percentage error)

Regression Metrics

Equal weighting of errors	MAE	$\frac{1}{n}\sum_{i=0}^{n-1} \lvert y_i - \hat{y}_i \rvert$
	MAPE	$\frac{100}{n}\sum_{i=0}^{n-1} \frac{\lvert y_i - \hat{y}_i \rvert}{y_i}$
Unequal weighting of errors	MSE	$\frac{1}{n}\sum_{i=0}^{n-1} \left(y_i - \hat{y}_i\right)^2$
	RMSE	$\sqrt{\frac{1}{n}\sum_{i=0}^{n-1} \left(y_i - \hat{y}_i\right)^2}$
	MSLE	$\frac{1}{n}\sum_{i=0}^{n-1} \left(\ln\left(1 = y_i\right) - \ln\left(1 + \hat{y}_i\right)\right)^2$
Data Variance	R^2	$1 - \frac{\sum_{i=0}^{n-1} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=0}^{n-1} \left(y_i - \bar{y}_i\right)^2}$
	Explained Var	$1 - \frac{Var\left(y - \hat{y}\right)}{Var\left(y\right)}$

Data Quality

Factors Which Affect Data Quality

Analysis is only ever as good as the data its built upon.

Factors Which Affect Data Quality

Analysis is only ever as good as the data its built upon.

- Data Definition
- Data Collection
- Data Processing
- Data Representation

How Can Data Be Wrong

- Incomplete
- Inconsistent
- Inaccurate

Processes to Assure Data Quality

- Data Provenance
- Sanity Checks
- Exploratory Data Analysis