Digital Transformation of Healthcare

Evaluating Predictions & Data Quality

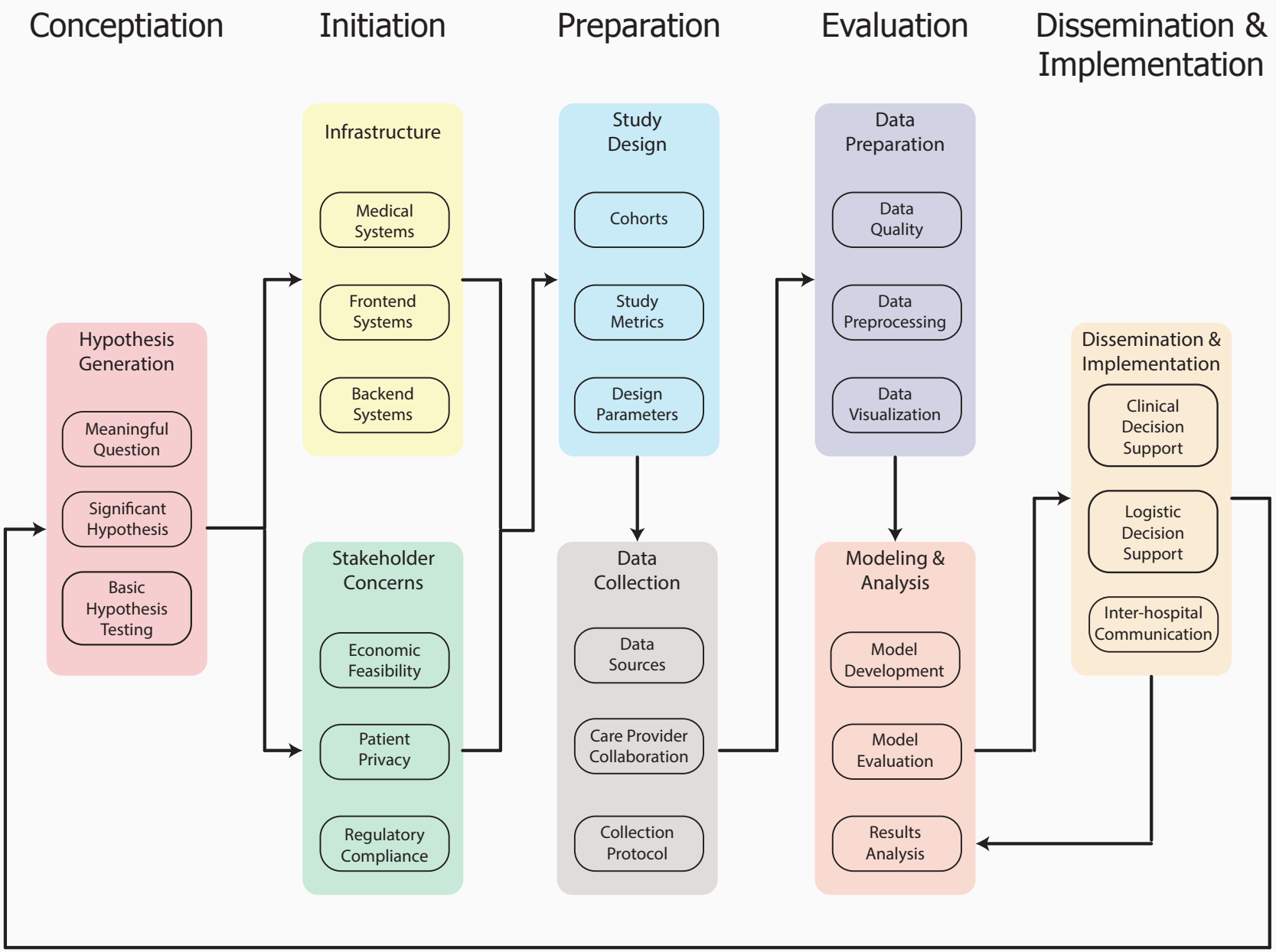
Michoel Snow, MD PhD, Glen Ferguson, PhD

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


Conceptiation
Initiation
Preparation
Evaluation
Dissemination & Implementation
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
Dissemination & Implementation
Clinical Decision Support
Logistic Decision Support
Inter-hospital Communication

Metrics for Evaluation of Classification Models

Terms and Questions

Terms

- Accuracy
- Specificity
- Sensitivity
- Positive Predictive Value
- Negative Predictive Value
- Likelihood Ratio
- ROC & AUC
- F1 Score

Questions

Terms and Questions

Terms

- Accuracy
- Specificity
- Sensitivity
- Positive Predictive Value
- Negative Predictive Value
- Likelihood Ratio
- ROC & AUC
- F1 Score

Questions

- Is accuracy a useful metric?
- What information is conveyed by sensitivity vs specificity ?
- What information do the PPV and NPV add?
- Intuitively, how do sensitivity, specificity, likelihood ratios and ROC connect?
- Is the F1 score a more robust metric than the ROC and AUC?

Clinical Cases

- Low dose CT for detecting lung cancer (LDCT)[1]
- Ultrasound detection of abdominal aortic aneurysms (AAA)[2]
- Blood pressure monitoring in adolescents using home machines (HTN)[3]
- Detecting suicidality among adolescent outpatients by clinicians versus trained raters using the Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS-PL) [4]

---

[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.

[4] Holi, M. M., Pelkonen, M., Karlsson, L., Tuisku, V., Kiviruusu, O., Ruuttu, T., & Marttunen, M. (2008). Detecting suicidality among adolescent outpatients: evaluation of trained clinicians' suicidality assessment against a structured diagnostic assessment made by trained raters. BMC psychiatry, 8(1), 97.

Confusion Matrix

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

HTN

Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

Confusion Matrix

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

HTN

Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

- What is the accuracy of these tests?

Confusion Matrix

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

HTN

Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

- What is the accuracy of these tests?
- Are these good screening tests and/or diagnostic tests?

Confusion Matrix

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

HTN

Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

- What is the accuracy of these tests?
- Are these good screening tests and/or diagnostic tests?
- How does the PPV and NPV affect your opinion of their utility?

Confusion Matrix

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

HTN

Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

- What is the accuracy of these tests?
- Are these good screening tests and/or diagnostic tests?
- How does the PPV and NPV affect your opinion of their utility?
- When are sensitivity, specificity, PPV and NPV appropriate tests?

Confusion Matrix

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

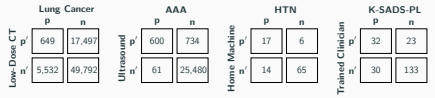
HTN

Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

- What is the accuracy of these tests?
- Are these good screening tests and/or diagnostic tests?
- How does the PPV and NPV affect your opinion of their utility?
- When are sensitivity, specificity, PPV and NPV appropriate tests?
- How are sensitivity, specificity, PPV and NPV affected by prevalence?



- What is the accuracy of these tests?
- Are these good screening tests and/or diagnostic tests?
- How does the PPV and NPV affect your opinion of their utility?
- When are sensitivity, specificity, PPV and NPV appropriate tests?
- How are sensitivity, specificity, PPV and NPV affected by prevalence?

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

Case	**TP**	**FP**	**TN**	**FN**	**Sens**	**Spec**	**PPV**	**NPV**	**Acc**	**F1**
LDCT	649	17,497	49,792	5,532	10	74	4	90	69	6
AAA	600	734	25,480	61	91	97	45	100	97	60
HTN	17	6	65	14	55	92	74	82	80	63
KSADS	32	23	133	30	52	85	58	82	76	55

Combined Statistics

Lung Cancer

Low-Dose CT	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound	p	n
p′	600	734
n′	61	25,480

HTN

Home Machine	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician	p	n
p′	32	23
n′	30	133

- What are 4 'sensible' pairings of the base stats

Combined Statistics

Lung Cancer

Low-Dose CT

	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound

	p	n
p′	600	734
n′	61	25,480

HTN

Home Machine

	p	n
p′	17	6
n′	14	65

K-SADS-PL

Trained Clinician

	p	n
p′	32	23
n′	30	133

- What are 4 'sensible' pairings of the base stats
- What are the different ways to combine the base stats into summary statistics (hint: what are the basic ways to combine any numbers)?
  - Work through each of the four clinical cases

Combined Statistics

Lung Cancer

Low-Dose CT

	p	n
p′	649	17,497
n′	5,532	49,792

AAA

Ultrasound

	p	n
p′	600	734
n′	61	25,480

HTN

Home Machine

	p	n
p′	17	6
n′	14	65

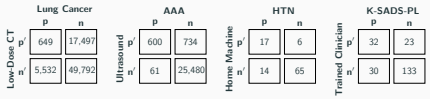
K-SADS-PL

Trained Clinician

	p	n
p′	32	23
n′	30	133

- What are 4 'sensible' pairings of the base stats
- What are the different ways to combine the base stats into summary statistics (hint: what are the basic ways to combine any numbers)?
  - Work through each of the four clinical cases
- What determines the split of positive cases into TP vs FN and negative cases into TN vs FP?

Combined Statistics

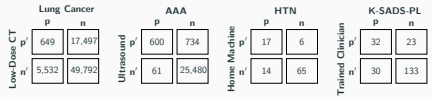


	Lung Cancer		AAA		HTN		K-SADS-PL	
	p	n	p	n	p	n	p	n
Low-Dose CT p′	649	17,497						
Low-Dose CT n′	5,532	49,792						
Ultrasound p′			600	734				
Ultrasound n′			61	25,480				
Home Machine p′					17	6		
Home Machine n′					14	65		
Trained Clinician p′							32	23
Trained Clinician n′							30	133

- What are 4 'sensible' pairings of the base stats
- What are the different ways to combine the base stats into summary statistics (hint: what are the basic ways to combine any numbers)?
  - Work through each of the four clinical cases
- What determines the split of positive cases into TP vs FN and negative cases into TN vs FP?

- sens/spec, PPV/NPV, sens/PPV, spec/NPV
- the 4 basic operations are add, subtract, multiply and divide **add Ex**
- **Add** adding just gives you the numbers themselves without an idea of how they each contribute. averaging sens/spec or ppv/npv is a good summary of how they perform. Averaging sens/ppv tells you how well you can predict TP taking into account the reliability of the test and the prevalence of the disease, and specifically ignoring the effects of TN.
- **Subtract** Not really a helpful metric as the difference between values doesn't tell you much about the values themselves. F1 score combines the typical mean and the difference between the values
- **multiply** Similar to averaging and F1 but punishes if both lower values
- **divide** dividing two probabilities gives you an odds ratio, i.e., how much more likely the numerator is to happen than the denominator.

Combined Statistics



Lung Cancer		P	n
Low-Dose CT	p′	649	17,497
	n′	5,532	49,792

AAA		P	n
Ultrasound	p′	600	734
	n′	61	25,480

HTN		P	n
Home Machine	p′	17	6
	n′	14	65

K-SADS-PL		P	n
Trained Clinician	p′	32	23
	n′	30	133

- What are 4 'sensible' pairings of the base stats
- What are the different ways to combine the base stats into summary statistics (hint: what are the basic ways to combine any numbers)?
  - Work through each of the four clinical cases
- What determines the split of positive cases into TP vs FN and negative cases into TN vs FP?

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

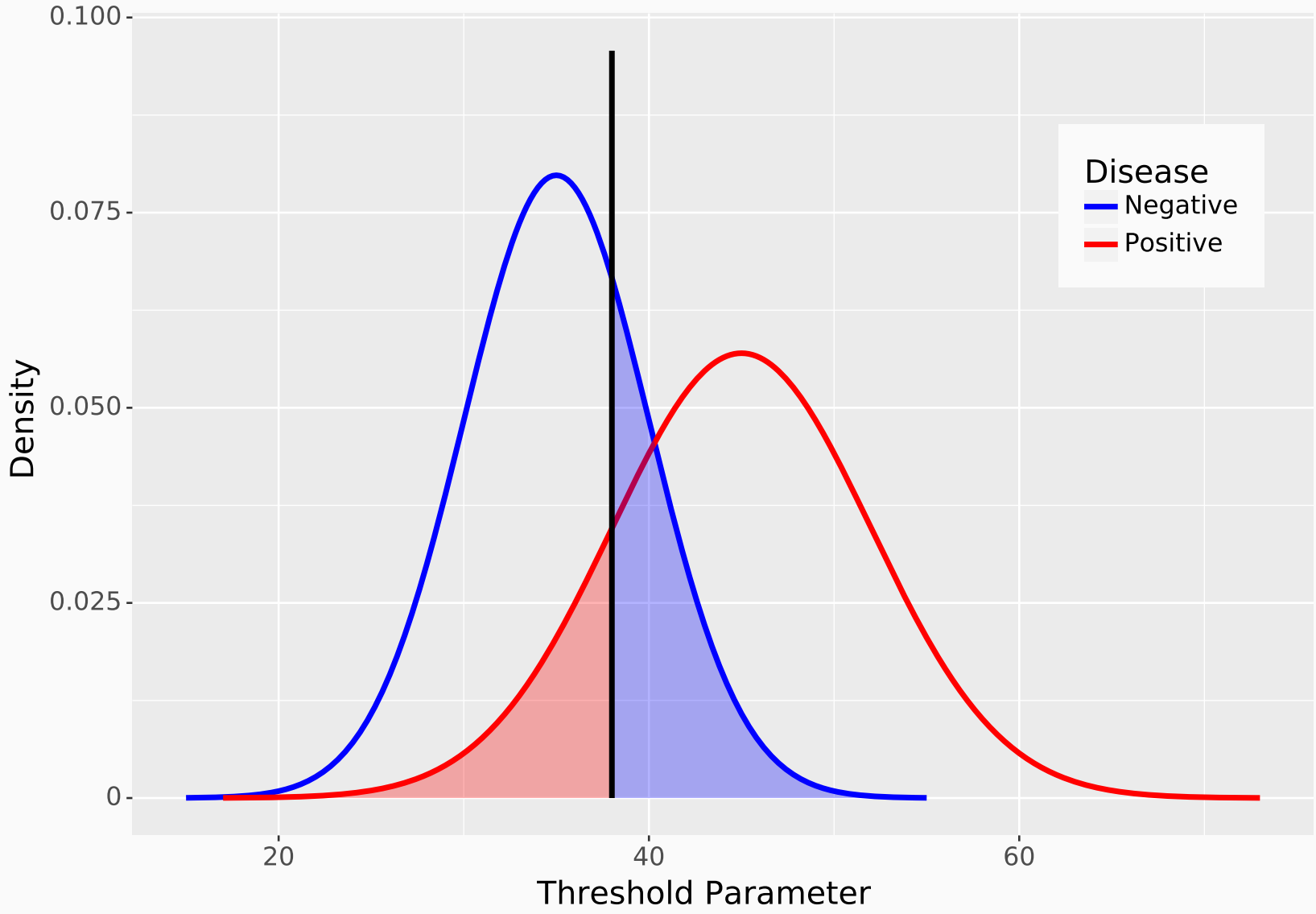
$$LR- = \frac{1 - sensitivity}{specificity} = \frac{P(T- \mid D+)}{P(T- \mid D-)}$$

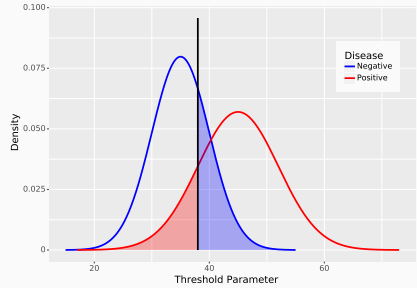
Likelihood Ratio	Approximate Change in Probability(%)
0.1	-45
0.2	-30
0.5	-15
1	0
2	+15
5	+30
10	+45

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

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

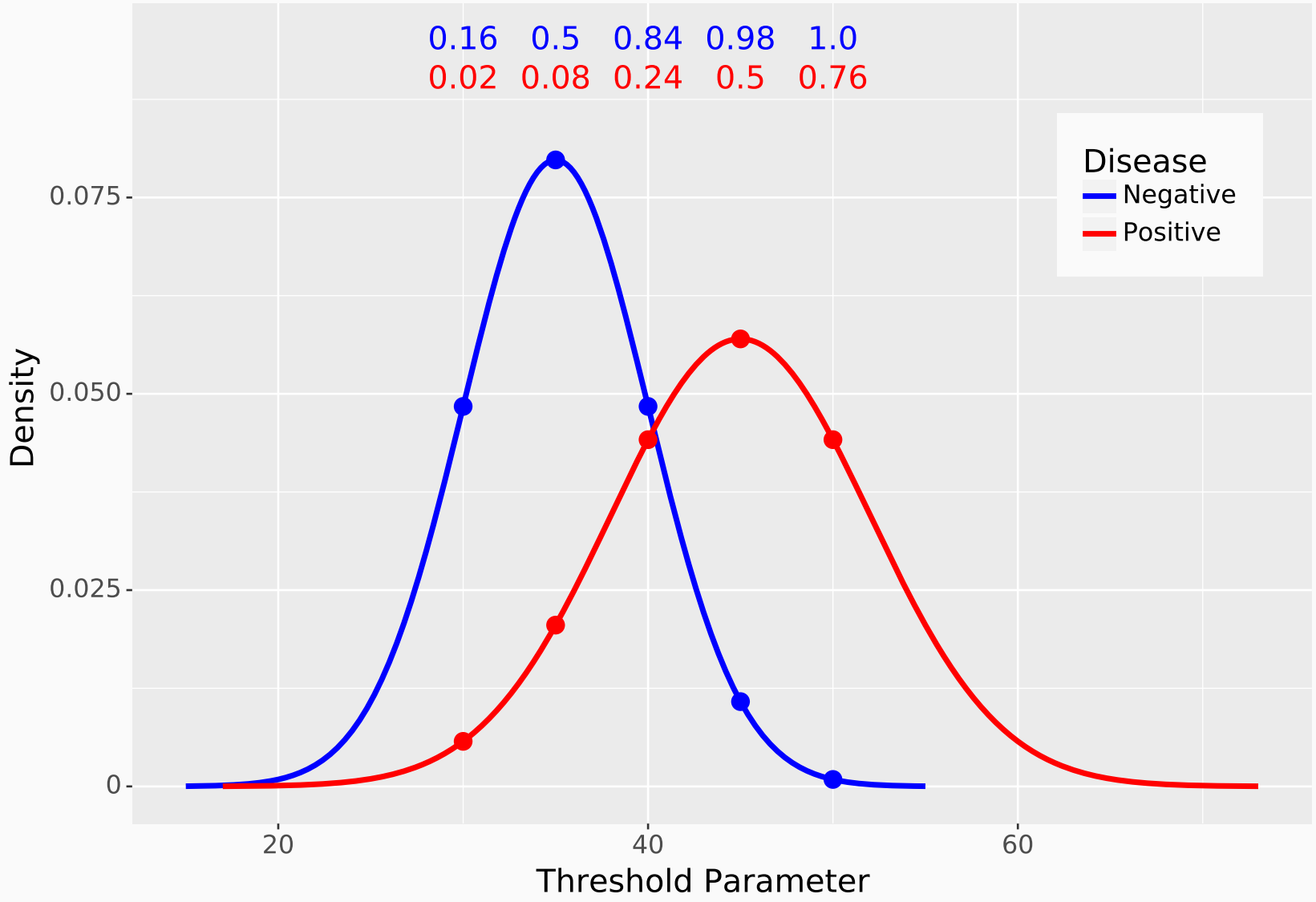
Hypothesis Testing

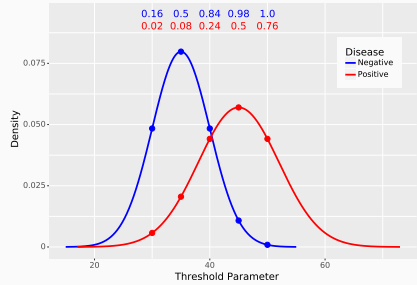




- these two curves represent the positive and negative cases
- As the threshold parameter from right to left, your sensitivity increases but you specificity decreases
- In order to calculate specific values we need to know what the areas under the curve are, for different thresholds

Hypothesis Testing





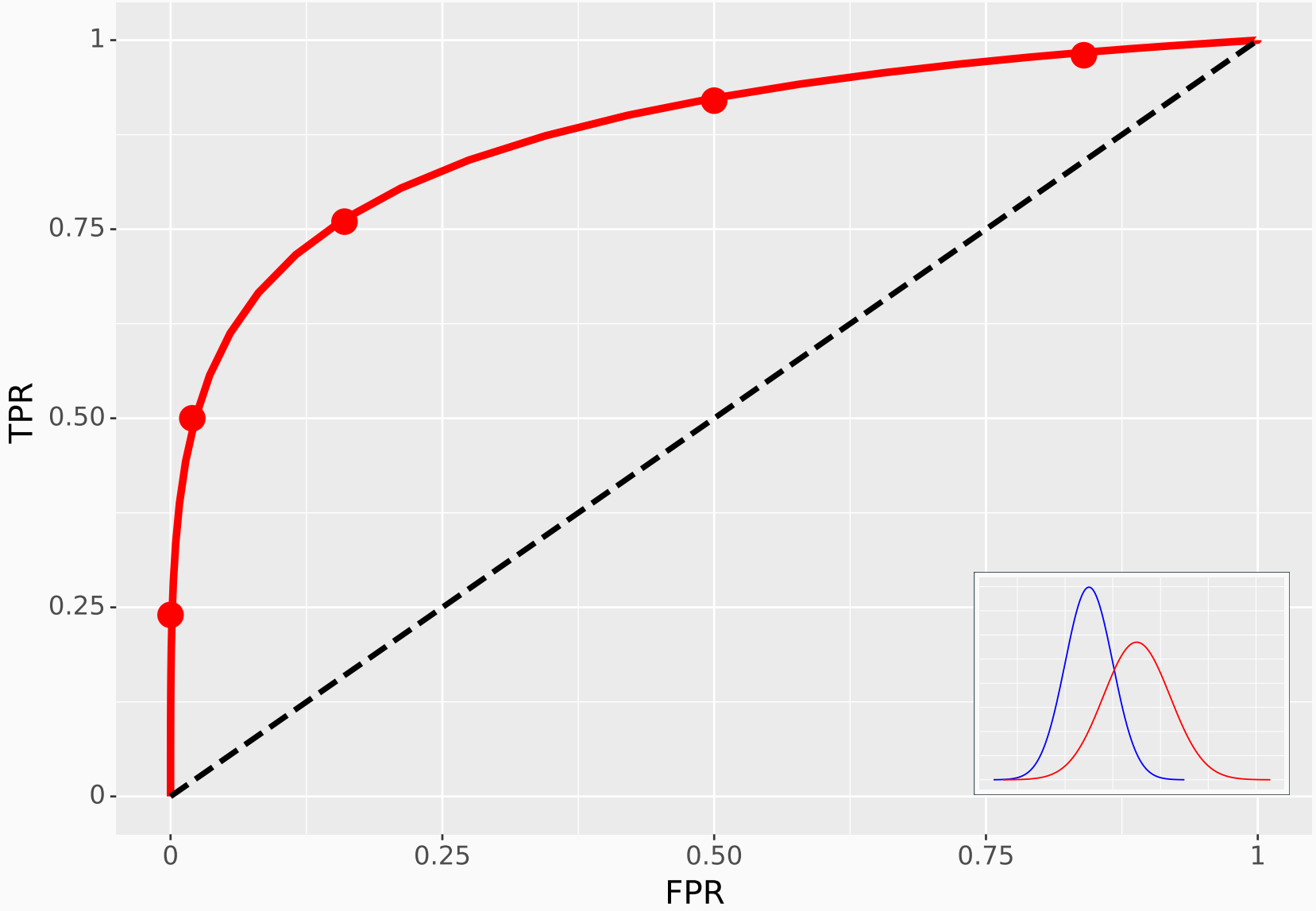
- Let's calculate the odds ratio for various points along the curve

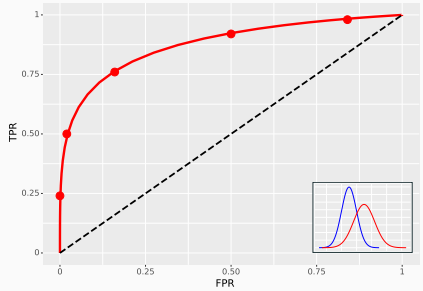
N/P	Sens	Spec	Odds	PPV	NPV	F1
0.16/002	0.98	0.16	1.17	0.54	0.89	0.70
0.5/0.08	0.92	0.50	1.84	0.65	0.86	0.76
0.84/0.24	0.76	0.84	4.78	0.83	0.78	0.79
0.98/0.5	0.50	0.98	26.32	0.96	0.66	0.66
1.0/0.76	0.24	1.00	∞	1	0.57	0.39

N/P	PPV-25%P	NPV-25%P	PPV-75%P	NPV-75%P	f1-25%P	f1-75%P
0.16/002	0.28	0.96	0.78	0.73	0.44	0.87
0.5/0.08	0.38	0.95	0.85	0.68	0.54	0.88
0.84/0.24	0.61	0.91	0.93	0.54	0.68	0.84
0.98/0.5	0.89	0.86	0.99	0.40	0.64	0.66
1.0/0.76	1	0.80	1.00	1	0.31	0.39

- both of these curves have an area of 1, so we need to multiply each metric by the percentage of patients to see how prevalence affects the outcome

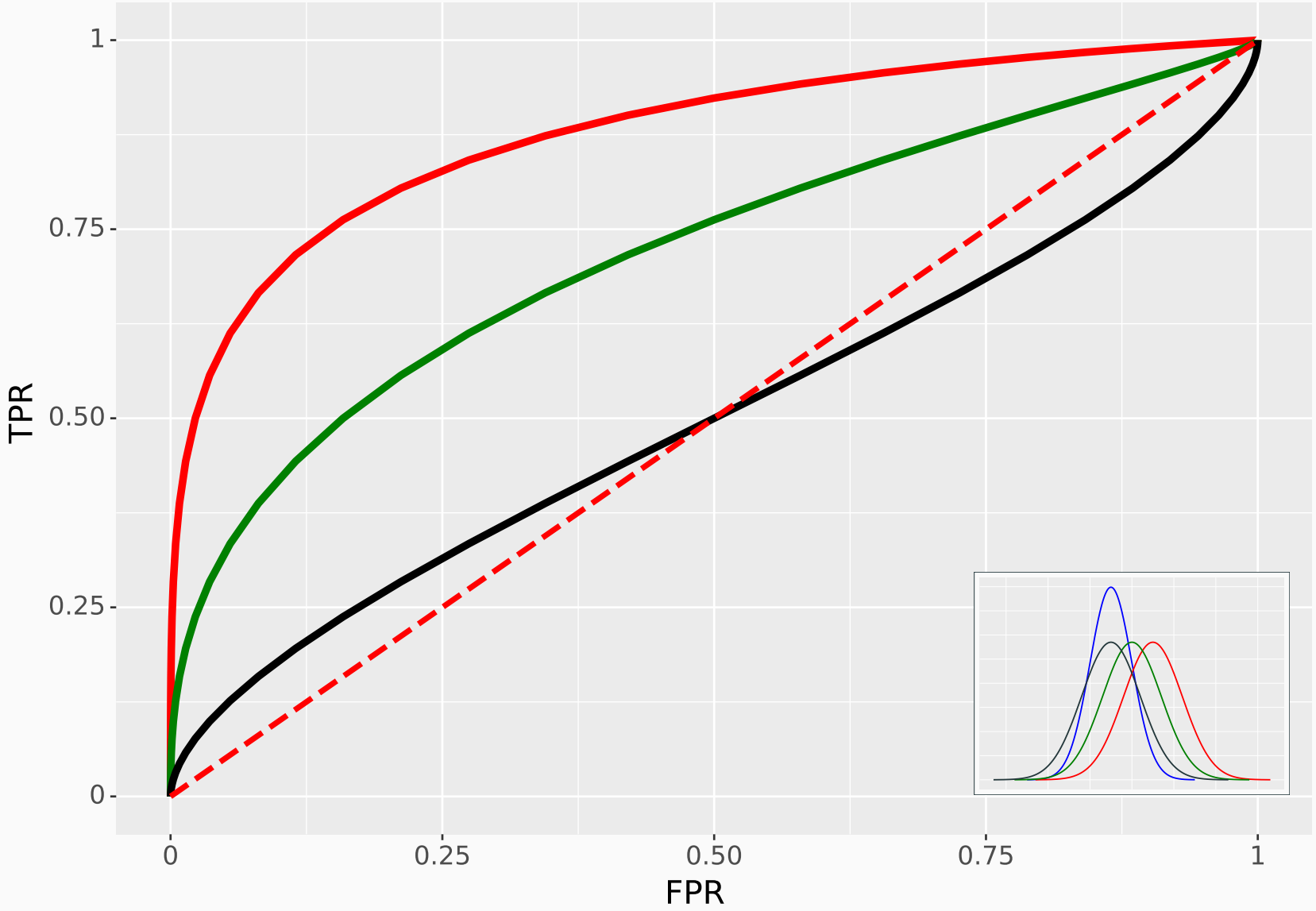
Receiver Operating Characteristic





1. What does the Area Under the Curve (AUC) correspond to? - Given a positive test result what are the chances that the subject is truly positive irrespective of prevalence?
2. What does the diagonal correspond to? - equal probabilities
3.
4. PPV is threshold dependent, while AUC is threshold independent but variable dependent

Receiver Operating Characteristic



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?

- 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	$1 - \frac{Var\,(y - \hat{y})}{Var\,(y)}$

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

Old Slides

Clinical Cases

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

Clinical Cases

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

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

Clinical Cases

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

Confusion Matrix

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

Confusion Matrix

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

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

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

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

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

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

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

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

Confusion Matrix

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

Condition	Stats	Example
High Sensitivity, Low Specificity	$\mathbf{p}' >> \mathbf{n}'$	test is always positive
Low Sensitivity, High Specificity	$\mathbf{n}' >> \mathbf{p}'$	test is always negative
High PPV, Low NPV	$\mathbf{p} >> \mathbf{n}$	high disease prevalence
Low PPV, High NPV	$\mathbf{n} >> \mathbf{p}$	low disease prevalence
High PPV, Low Sensitivity	$FN >> FP$	say they are negative most of the time for a high preva
High Sensitivity, Low NPV		
High Specificity, Low NPV		
High PPV, Low Specificity		

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

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

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

Likelihood Ratios

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

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

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

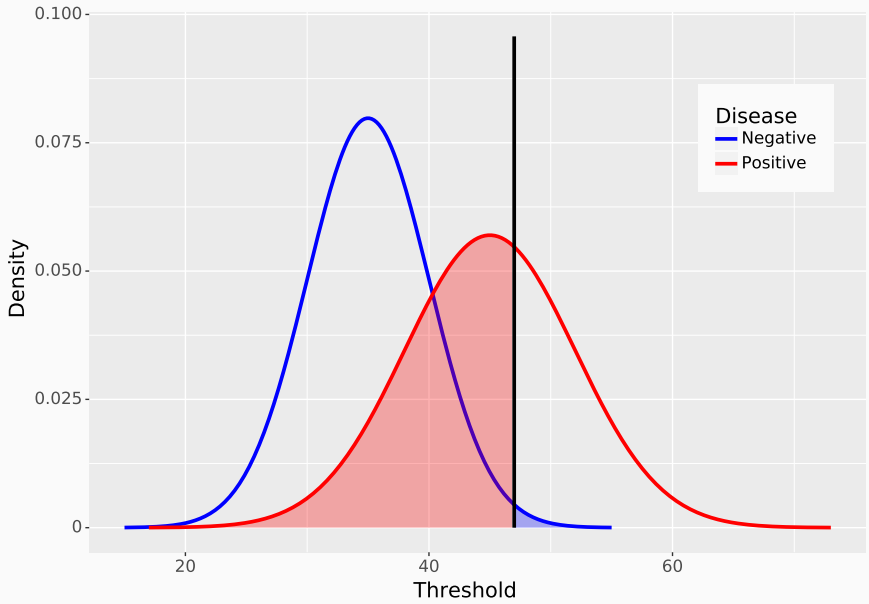
Likelihood Ratios

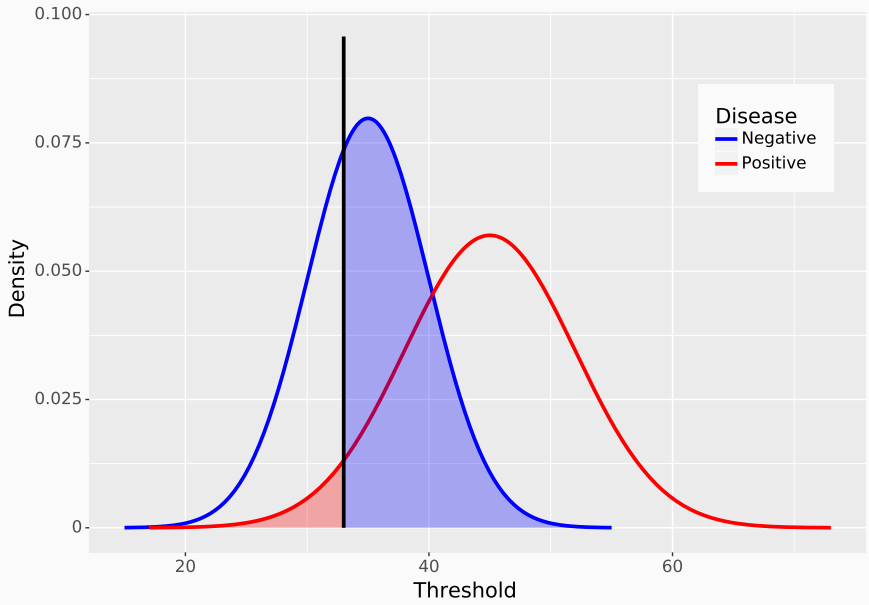
Likelihood Ratio	Approximate Change in Probability(%)
0.1	-45
0.2	-30
0.5	-15
1	0
2	+15
5	+30
10	+45

$$\text{Change in post test probability} \approx 0.2 \times \ln LR$$ [5]

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

Discrimination Thresholds





F1 Score

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

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

How does F1 differ from AUC?

F1 Score

predicted outcome \ actual outcome	p	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.