Definitions

- True Positive (TP) – observation is positive and was predicted as positive
- True Negative (TN) – observation is negative and was predicted as negative
- False Positive (FP) – observation is negative and was predicted as positive
- False Negative (FN) – observation is positive and was predicted as negative
- Classification accuracy – proportion of observations classified correctly = (#TP + #TN)/N

Evaluation

- Possible evaluation measures:
  - *Classification Accuracy* – or conversely, the *classification error rate* = 1 – classification accuracy
  - *Differential penalties* – different errors may have different costs
  - Area Under Receiver-Operating Characteristic (ROC) Curve (AUC)

- How reliable are the predicted results ?

Classifier error rate

- Natural performance measure for classification problems: *error rate (or success rate)*
  - *Success*: instance's class is predicted correctly
  - *Error*: instance's class is predicted incorrectly
  - *Error rate*: proportion of errors made over the whole set of instances
  - *Success rate*: proportion of errors made over the whole set of instances = 1 – Error rate

Cost-Benefit

- Cost-benefit is similar to classification error, except that we might want to count some errors to be worse than others
- Consider a disease where treatment involves taking a simple cheap and benign medicine to cure it (e.g. taking vitamin C), but if the disease is unchecked, the consequences are severe.
- In that case, I would want to put much more weight on missing positives than missing negatives; i.e. consequences of false negatives are much worse than false positives.
- For example, we might make the penalty (cost) for a false negative be 10 or 100 times that of a false positive

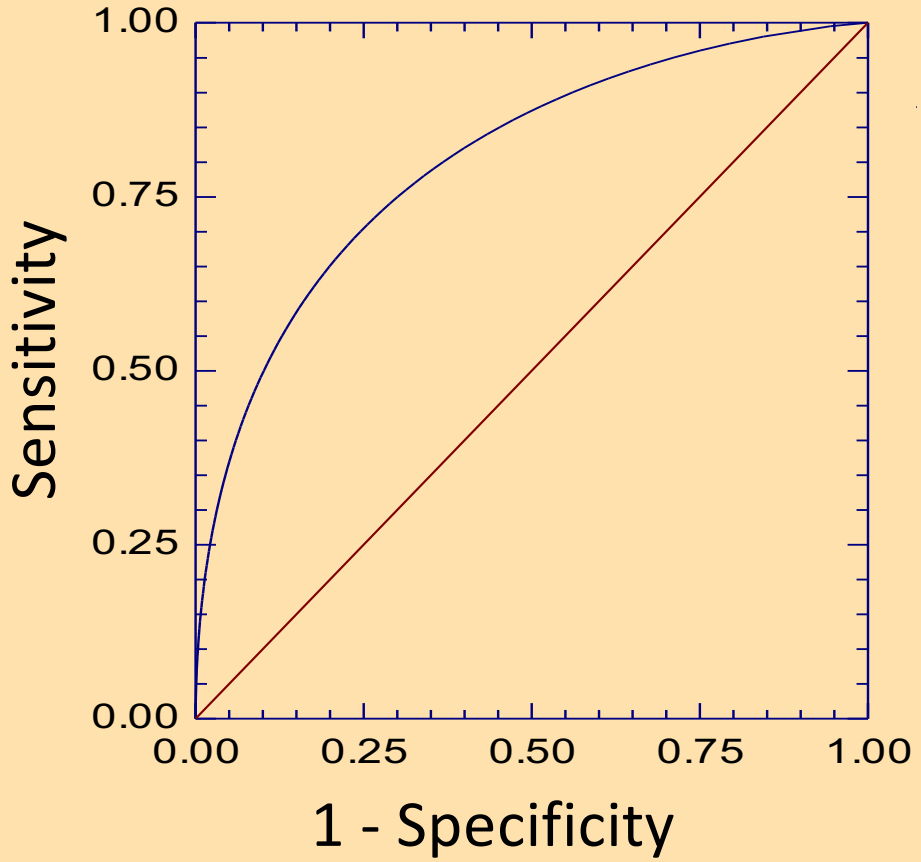
Sensitivity and Specificity

Test\Truth	Disease Free	Diseased	Total
Negative	TN	FN	FN+TN
Positive	FP	TP	TP+FP
Total	FP+TN	TP+FN	N

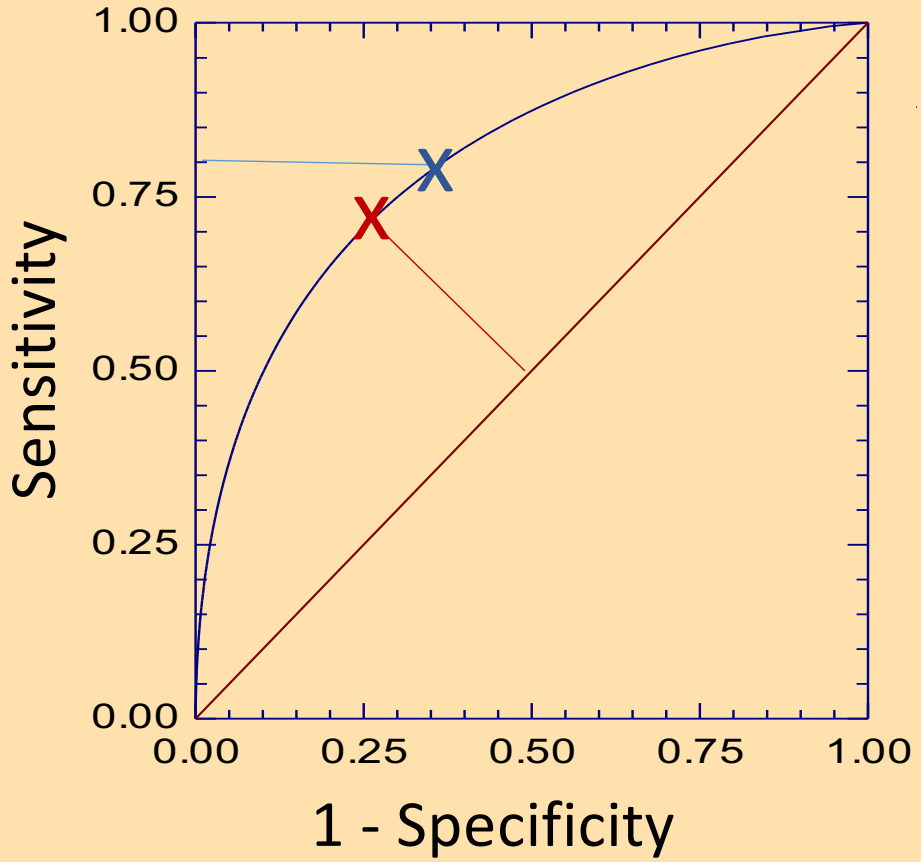
Sensitivity and Specificity

- $\text{Sensitivity} = \frac{\#\text{TP}}{\#\text{TP}+\#\text{FN}} =$
  Proportion of positives correctly identified as positive
- $\text{Specificity} = \frac{\#\text{TN}}{\#\text{TN}+\#\text{FP}} =$
  Proportion of negatives correctly identified as negative
- There is another trade-off here between sensitivity and specificity
  - Imagine a classification procedure that gives you a probability of being positive for a disease (such classifiers are common)
  - To classify individuals we need to choose a probability threshold to identify them as positive vs. negative
  - Increasing the threshold increases specificity but reduces sensitivity
  - Decreasing the threshold increases sensitivity but reduces specificity

Receiver operating characteristic (ROC) curve

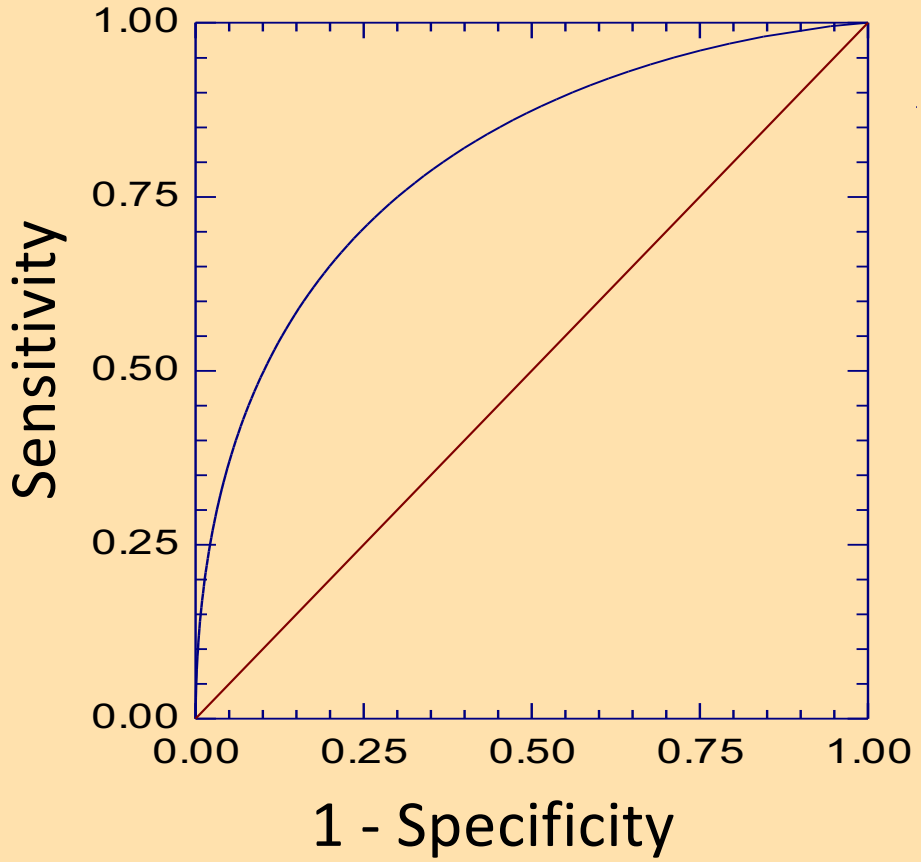


Choosing threshold based on ROC



- Many possible thresholds, each corresponding to one point on curve
- Choosing a single one depends on context
- Some studies say (for example) give me best specificity for a sensitivity of at least 0.8 (blue x)
- General alternative – maximal perpendicular from diagonal (red x)
- We will not be concerned with choosing thresholds, rather, we will use ROC to generate a metric

Area under the (ROC) curve (AUC)



- AUC gives a measure of how well the predictor does across all possible thresholds
- Ideal AUC is 1
- AUC for random predictor = 0.5
- Objective is to find classifier that maximizes AUC

Error rate vs. AUC

- When should we choose error rate and when AUC?
- Error rate (and cost-benefit) work well when your dataset is representative of the larger population (metric incorporates prevalence – e.g. proportion of people in population that have disease)
- AUC works better when your dataset is not representative of larger population, e.g. in “case-control” studies (metric is independent of prevalence)