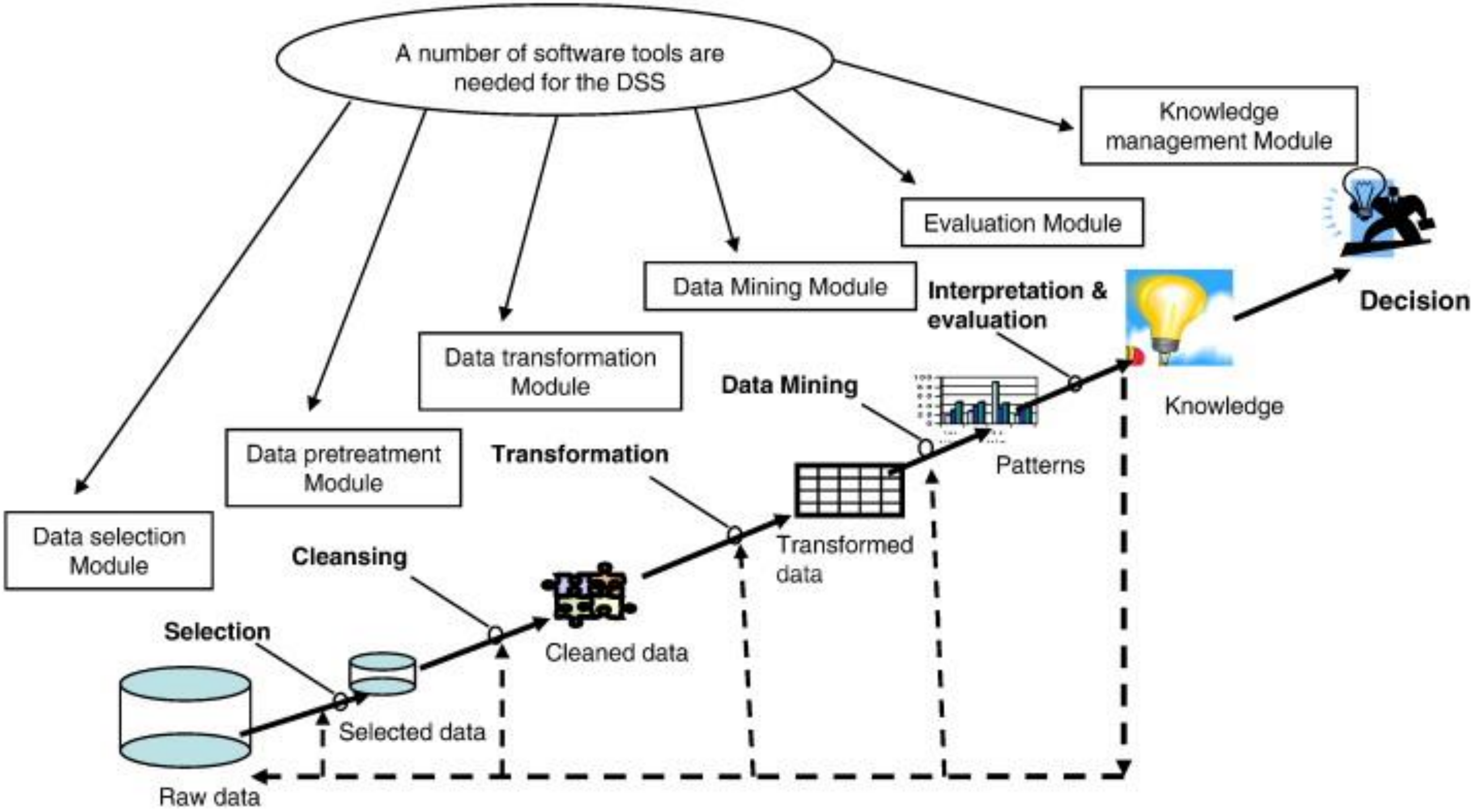
Lecture Notes

Big Data in Medical Informatics

Week 11:

Data Analytics in Medicine – Part 2

i5 | RWTHAACHEN UNIVERSITY



Ayed, B. M., Ltifi, H., Kolski,C. & Alimi, A. (2010) A usercentered approach for the design & implementation of KDD-based DSS: A case study in the healthcare domain. Decision Support Systems, 50, 64- 78.

Styles of Learning

Supervised

- Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

  $$Y = f(X)$$

- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.
- the focus is on accurate prediction

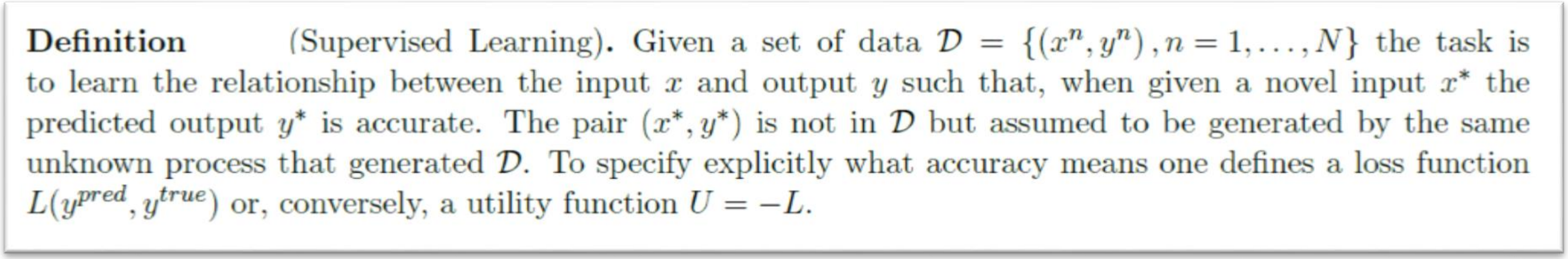
Unsupervised

- Unsupervised learning is where you only have input data (X) and no corresponding output variables.
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.
- to find compact descriptions of the data

Supervised Learning

- Aim : to discover underlying relationships between covariate variables (attributes, features)

- Dependent variable : outcome



Definition (Supervised Learning). Given a set of data $\mathcal{D} = \{(x^n, y^n), n = 1, \ldots, N\}$ the task is to learn the relationship between the input x and output y such that, when given a novel input x^* the predicted output y^* is accurate. The pair (x^*, y^*) is not in $\mathcal{D}$ but assumed to be generated by the same unknown process that generated $\mathcal{D}$. To specify explicitly what accuracy means one defines a loss function $L(y^{pred}, y^{true})$ or, conversely, a utility function $U = -L$.

- If the output is one of a discrete number of possible `classes', this is called a classification problem
  - generates class labels

- If the output is continuous, this is called a regression problem
  - Generates real valued outcomes

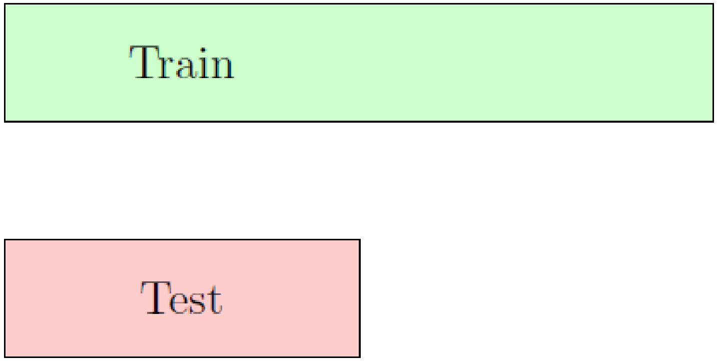
Supervised Learning in Clinical Prediction

Types of outcomes:

- Continuous outcomes: e.g. medical cost prediction
  - Linear regression
- Binary outcomes: e.g. diseases, death/alive
  - Clinical prediction models, logistic regression, binary classification trees, Bayesian models
- Categorical outcomes: Multiclass classification problem such as tumor classification, multiple disease diagnostic
  - Polytomous logistic regression, decision trees, ..
- Ordinal outcomes: grade/severity of disease
- Survival outcomes: predict time to event of interest

Supervised Learning

- In training and evaluating a model, conceptually there are two sources of data.



Train

Test

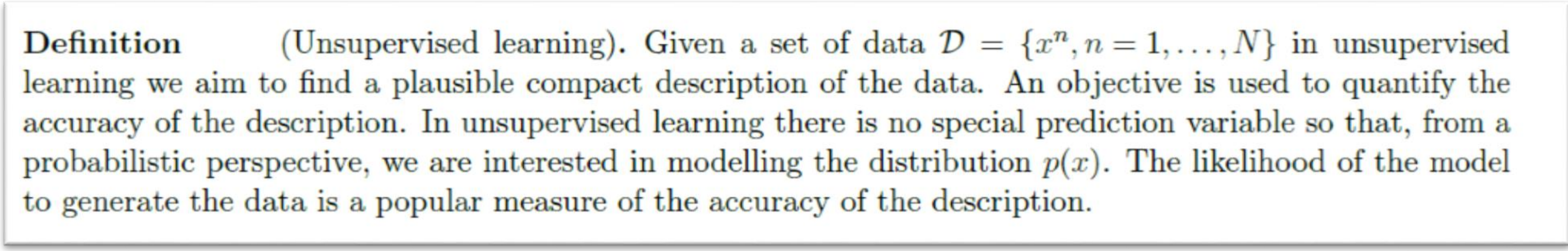
- The parameters of the model are set on the basis of the train data only.
- If the test data is generated from the same underlying process that generated the train data, an unbiased estimate of the generalisation performance can be obtained by measuring the test data performance of the trained model.
- Importantly, the test performance should not be used to adjust the model parameters since we would then no longer have an independent measure of the performance of the model.

Unsupervised Learning

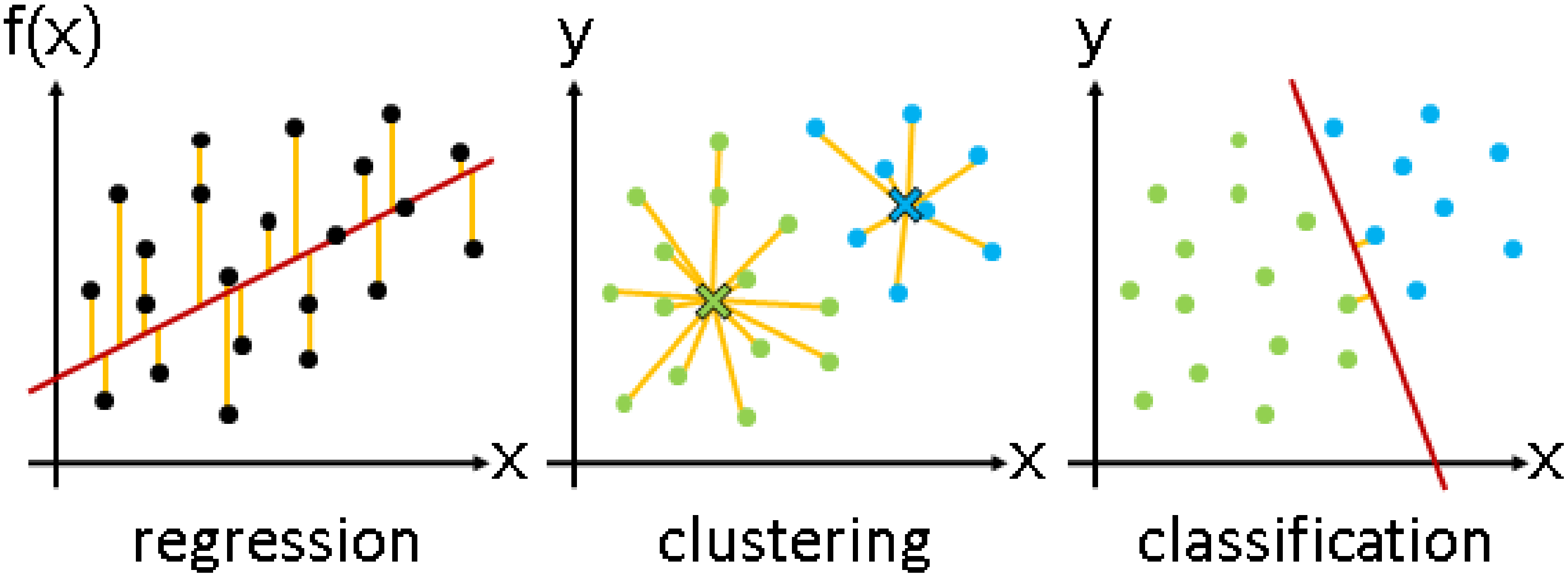
Unsupervised learning problems can be further grouped into clustering and association problems.

- Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

- Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.



Definition (Unsupervised learning). Given a set of data $\mathcal{D} = \{x^n, n = 1, \ldots, N\}$ in unsupervised learning we aim to find a plausible compact description of the data. An objective is used to quantify the accuracy of the description. In unsupervised learning there is no special prediction variable so that, from a probabilistic perspective, we are interested in modelling the distribution $p(x)$. The likelihood of the model to generate the data is a popular measure of the accuracy of the description.


f(x)
x
regression
y
x
clustering
y
x
classification

Naïve Bayes Classifier

- Bayesian statistics allow one to make an estimate about the likelihood of a claim and then update these estimates as new evidence becomes available.
- The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors.
- A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets.
- Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Probability

- Rules of Probability for Discrete Variables:

The probability p(x = x) of variable x being in state x is represented by a value between 0 and 1

p(x = x) = 1 means that we are certain x is in state x.

p(x = x) = 0 means that we are certain x is not in state x.

Values between 0 and 1 represent the degree of certainty of state occupancy.

The summation of the probability over all the states is 1 (called the normalisation condition):

$$\sum_{\mathrm{x} \in \mathrm{dom}(x)} p(x = \mathrm{x}) = 1$$

Probability

- Two variables x and y can interact through

p(x = a or y = b) = p(x = a) + p(y = b) - p(x = a and y = b)

Or, more generally, we can write

p(x or y) = p(x) + p(y) - p(x and y)

We can use the shorthand p(x, y) for p(x and y).

p(y, x) = p(x, y) and p(x or y) = p(y or x)

Probability

- Independence:

Variables x and y are independent if knowing the state (or value in the continuous case) of one variable gives no extra information about the other variable

$p(x, y) = p(x)p(y)$

Provided that $p(x) \neq 0$ and $p(y) \neq 0$ independence of x and y is equivalent to

$p(x|y) = p(x) \leftrightarrow p(y|x) = p(y)$

If $p(x|y) = p(x)$ for all states of x and y, then the variables x and y are said to be independent

Probability

- Example:

x denote the day of the week in which females are born,
y denote the day in which males are born,
dom(x) = dom(y) = {1, …, 7}.

It is reasonable to expect that x is independent of y.

We randomly select a woman from the phone book, Alice, and find out that she was born on a Tuesday.

We also randomly select a male at random, Bob. Before phoning Bob and asking him, what does knowing Alice's birth day add to which day we think Bob is born on?

Under the independence assumption, the answer is nothing.
It means that knowing when Alice was born doesn't provide any extra information than we already knew about Bob's birthday, $p(y|x) = p(y)$.

Probability

Conditional Probability / Bayes' Rule)

- The probability of event x conditioned on knowing event y (or more shortly, the probability of x given y) is defined as

$$p(x|y) \equiv \frac{p(x,y)}{p(y)}$$

conditional independence assumption: We assume that attribute values are independent of each other given the class

since p(x, y) = p(y, x)

p(x|y)= p(x) p(y) / p(y) ---------- since p(y) = p(y|x)

p(x|y)= p(x) p(y|x) / p(y)

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Bayes Rule

- Prior, conditional and joint probability for random variables
  - Prior probability: $P(x)$
  - Conditional probability: $P(x_1 \mid x_2), P(x_2 \mid x_1)$
  - Joint probability: $\mathbf{x} = (x_1, x_2), P(\mathbf{x}) = P(x_1, x_2)$
  - Conditional probability:

$$P(x_1, x_2) = P(x_2 \mid x_1)P(x_1) = P(x_1 \mid x_2)P(x_2)$$

  - Independence:

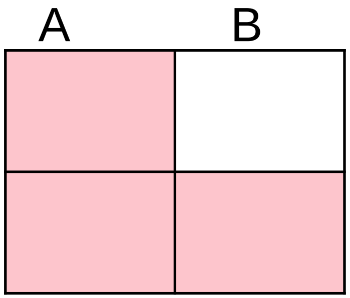
$$P(x_2 \mid x_1) = P(x_2), P(x_1 \mid x_2) = P(x_1), P(x_1, x_2) = P(x_1)P(x_2)$$

- Bayesian Rule

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c)P(c)}{P(\mathbf{x})}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Understanding the theory



- We have two cups : A and B
- Cup A contains 20 red balls
- Cup B contains 10 red balls and 10 white balls

- Alice draws a ball from one of the cups.
- We know that ball is red .
- Question: which cup did she draw ?

- Hypothesis: she draws from B

- P(H|E) = P(H) P(E|H) / P(E)

What happens if there is no red ball in A ?

Bayes Classifiers

- *Find out the probability of the previously unseen instance belonging to each class, then simply pick the most probable class.*

- Bayesian classifiers use **Bayes theorem**, which says

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

- *p*(*cj* | *d*) = probability of instance *d* being in class *cj*,

  This is what we are trying to compute

- *p*(*d* | c*j*) = probability of generating instance *d* given class *cj*,

  We can imagine that being in class c*j*, causes you to have feature *d* with some probability

- *p(cj)*= probability of occurrence of class *cj*,

  This is just how frequent the class c*j*, is in our database

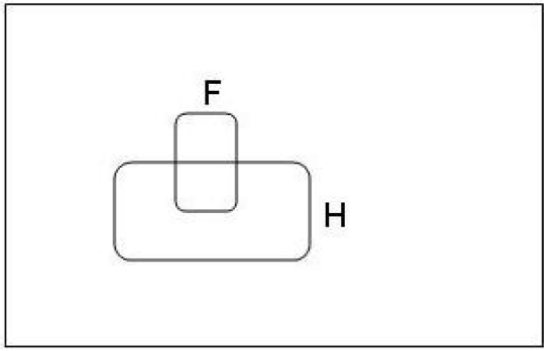
- *p(d)*= probability of instance *d* occurring

  Normalization factor

Bayes Classifiers

- S = “Have a headache”
- D = “Flu”

- P(S) = 1/10
- P(D) = 1/40
- P(S|D) = 1/2



- P(D|S) is the probability of the disease given the symptom. This is what we wish to infer.
- P(D) is the probability of the disease (within a population) this is a measurable quantity.
- P(S|D) is the probability of the symptom given the disease. We can measure this from the case histories of the disease.
- P(S) is the probability of the symptom in the population.

- P(D|S) = P(S|D) P(D) / P(S)= (½ * 1/40) / (1/10) = 1/8

Probabilistic Classification

- **M**aximum **A** **P**osterior (**MAP**) classification rule
  - For an input *x*, find the largest one from L probabilities output by a discriminative probabilistic classifier $P(c_1 \mid \mathbf{x}), \ldots, P(c_L \mid \mathbf{x})$.
  - Assign x to label c^* if $P(c^* \mid \mathbf{x})$ is the largest.
- Generative classification with the MAP rule
  - Apply Bayesian rule to convert them into posterior probabilities

$$P(c_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_i)P(c_i)}{P(\mathbf{x})} \propto P(\mathbf{x} \mid c_i)P(c_i)$$

$$\text{for } i = 1, 2, \cdots, L$$

Common factor for all L probabilities

  - Then apply the MAP rule to assign a label

Naïve Bayesian Classification

- Bayes classification

$$P(c\,|\,\mathbf{x}) \propto P(\mathbf{x}\,|\,c)P(c) = P(x_1,\cdots,x_n\,|\,c)P(c) \;\; \text{for } c = c_1,\ldots,c_L .$$

Difficulty: learning the joint probability $P(x_1,\cdots,x_n\,|\,c)$ is infeasible!

- Naïve Bayes classification
  - Assume all input features are class conditionally independent!

$$\begin{aligned} P(x_1,x_2,\cdots,x_n\,|\,c) &= \underline{P(x_1\,|\,x_2,\cdots,x_n,c)}P(x_2,\cdots,x_n\,|\,c) \\ &= \underline{P(x_1\,|\,c)}P(x_2,\cdots,x_n\,|\,c) \\ &= P(x_1\,|\,c)P(x_2\,|\,c)\cdots P(x_n\,|\,c) \end{aligned}$$

Applying the independence assumption

  - Apply the MAP classification rule: assign $\mathbf{x}' = (a_1,a_2,\cdots,a_n)$ to c^* if

$$\underbrace{[P(a_1\,|\,c^*)\cdots P(a_n\,|\,c^*)]}_{\text{estimate of } P(a_1,\cdots,a_n\,|\,c^*)}P(c^*) > \underbrace{[P(a_1\,|\,c)\cdots P(a_n\,|\,c)]}_{\text{esitmate of } P(a_1,\cdots,a_n\,|\,c)}P(c), \quad c \neq c^*, c = c_1,\cdots,c_L$$

Naïve Bayesian Classification

- Algorithm: Discrete-Valued Features
  - Learning Phase: Given a training set **S** of F features and L classes,

```
For each target value of $c_i$ $(c_i = c_1, \cdots, c_L)$
    $\hat{P}(c_i) \leftarrow$ estimate $P(c_i)$ with examples in S;
    For every feature value $x_{jk}$ of each feature $x_j$ $(j = 1, \cdots, F; k = 1, \cdots, N_j)$
        $\hat{P}(x_j = x_{jk} | c_i) \leftarrow$ estimate $P(x_{jk} | c_i)$ with examples in S;
```

Output: $F * L$ conditional probabilistic (generative) models

  - Test Phase: Given an unknown instance $\mathbf{x}' = (a'_1, \cdots, a'_n)$

“Look up tables” to assign the label c^* to **X’** if

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_n | c^*)]\hat{P}(c^*) > [\hat{P}(a'_1 | c_i) \cdots \hat{P}(a'_n | c_i)]\hat{P}(c_i), \quad c_i \neq c^*, c_i = c_1, \cdots, c_L$$

Bayes Classifiers

- Given than a person is smoker what is the probability of being Lung cancer ?

$$\mathrm{P(Lung\ Cancer\ |Smoker)} = \frac{\mathrm{P(Lung\ Cancer\ |Smoker)\ P(Smoker)}}{\mathrm{P(Lung\ Cnacer)}}$$

- P (Lung Cancer) = 3/8
- P(Smoker)= 5/8
- P(Smoker| Lung Cancer) = 2/5

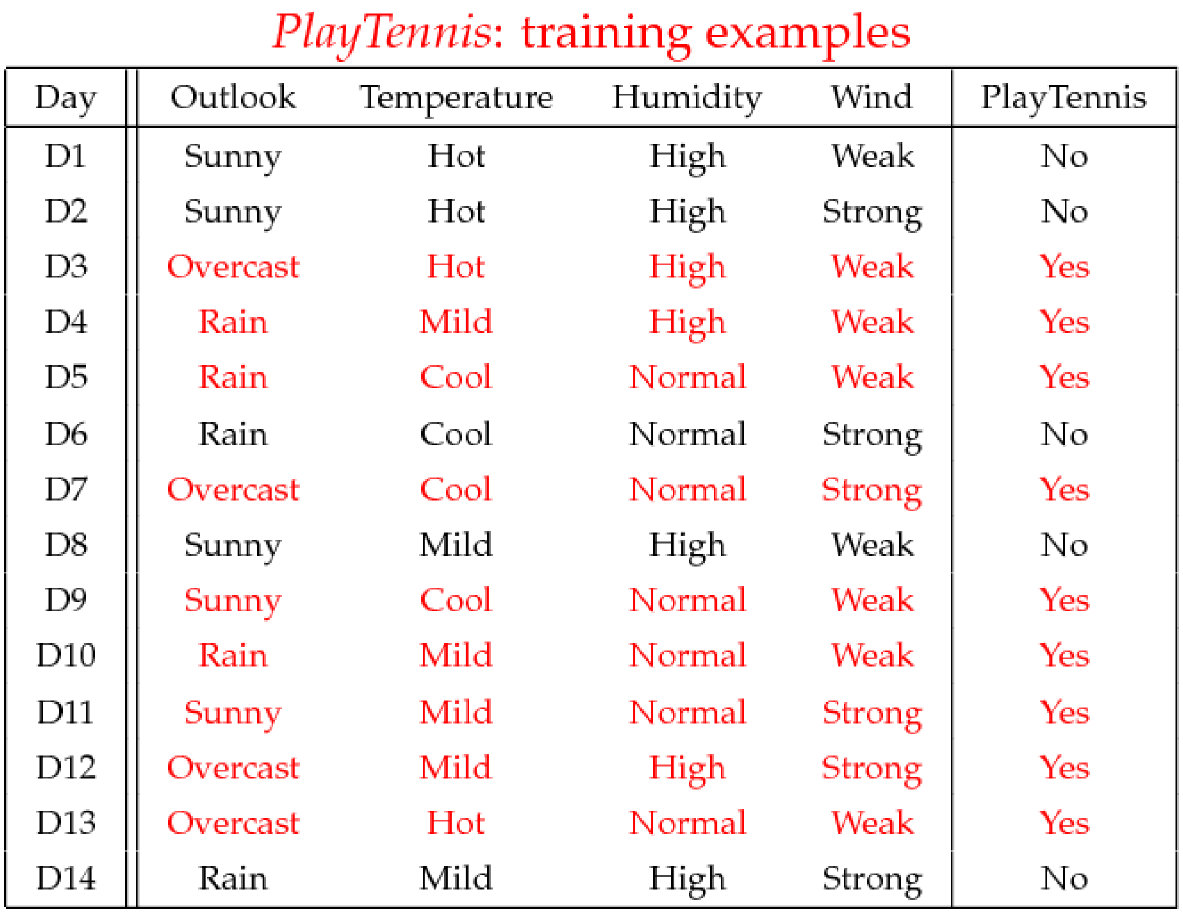
P(Lung Cancer | Smoker) = 0.25

Similarly calculate
P(Lung Cancer | Non- Smoker) = 0.125

Disease	Smoker
Lung Ca	Non-smoker
Flue	Smoker
Lung Ca	Smoker
Lung Ca	Smoker
Diabetes	Non-smoker
Dementia	Smoker
Flue	Smoker
Breast Ca	Non-smoker

Example

- Example: Play Tennis



PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=\textit{Yes}) = 9/14 \qquad P(\text{Play}=\textit{No}) = 5/14$

Example

- Test Phase
  - Given a new instance, predict its label

    $\mathbf{x}'=(\text{Outlook}=\textit{Sunny}, \text{Temperature}=\textit{Cool}, \text{Humidity}=\textit{High}, \text{Wind}=\textit{Strong})$

  - Look up tables achieved in the learning phrase

$P(\text{Outlook}=\textit{Sunny}|\text{Play}=\textit{Yes}) = 2/9$

$P(\text{Temperature}=\textit{Cool}|\text{Play}=\textit{Yes}) = 3/9$

$P(\text{Huminity}=\textit{High}|\text{Play}=\textit{Yes}) = 3/9$

$P(\text{Wind}=\textit{Strong}|\text{Play}=\textit{Yes}) = 3/9$

$P(\text{Play}=\textit{Yes}) = 9/14$

$P(\text{Outlook}=\textit{Sunny}|\text{Play}=\textit{No}) = 3/5$

$P(\text{Temperature}=\textit{Cool}|\text{Play}==\textit{No}) = 1/5$

$P(\text{Huminity}=\textit{High}|\text{Play}=\textit{No}) = 4/5$

$P(\text{Wind}=\textit{Strong}|\text{Play}=\textit{No}) = 3/5$

$P(\text{Play}=\textit{No}) = 5/14$

  - Decision making with the MAP rule

$P(\textit{Yes}|\mathbf{x}') \approx [P(\textit{Sunny}|\textit{Yes})P(\textit{Cool}|\textit{Yes})P(\textit{High}|\textit{Yes})P(\textit{Strong}|\textit{Yes})]P(\text{Play}=\textit{Yes}) = 0.0053$

$P(\textit{No}|\mathbf{x}') \approx [P(\textit{Sunny}|\text{No})\ P(\textit{Cool}|\text{No})P(\textit{High}|\textit{No})P(\textit{Strong}|\textit{No})]P(\text{Play}=\textit{No}) = 0.0206$

Given the fact $P(\textit{Yes}|\mathbf{x}') < P(\textit{No}|\mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

Naïve Bayes

- Algorithm: Continuous-valued Features
  - Numberless values taken by a continuous-valued feature
  - Conditional probability often modeled with the normal distribution

$$\hat{P}(x_j \mid c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji} : mean (avearage) of feature values x_j of examples for which $\mathrm{c} = c_i$

σ_{ji} : standard deviation of feature values x_j of examples for which $c = c_i$

  - Learning Phase:

    Output: normal distributions
  - Test Phase: Given an unknown instance
    - Instead of looking-up tables, calculate conditional probabilities with all the normal distributions achieved in the learning phrase
    - Apply the MAP rule to assign a label (the same as done for the discrete case)

Naïve Bayes

- Example: Continuous-valued Features
  - Temperature is naturally of continuous value.

    Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

    No: 27.3, 30.1, 17.4, 29.5, 15.1
  - Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$
$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

  - **Learning Phase**: output two Gaussian models for P(temp|C)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2 \times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2 \times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

Zero conditional probability

- If no example contains the feature value
  - In this circumstance, we face a zero conditional probability problem during test

    $$\hat{P}(x_1|c_i)\cdots\hat{P}(a_{jk}|c_i)\cdots\hat{P}(x_n|c_i)=0 \quad \text{for } x_j=a_{jk},\ \hat{P}(a_{jk}|c_i)=0$$

  - For a remedy, class conditional probabilities re-estimated with

$$\hat{P}(a_{jk}|c_i)=\frac{n_c+mp}{n+m}$$

(m-estimate)

n_c : number of training examples for which $x_j=a_{jk}$ and $c=c_i$

n : number of training examples for which $c=c_i$

p : prior estimate (usually, $p=1/t$ for t possible values of x_j)

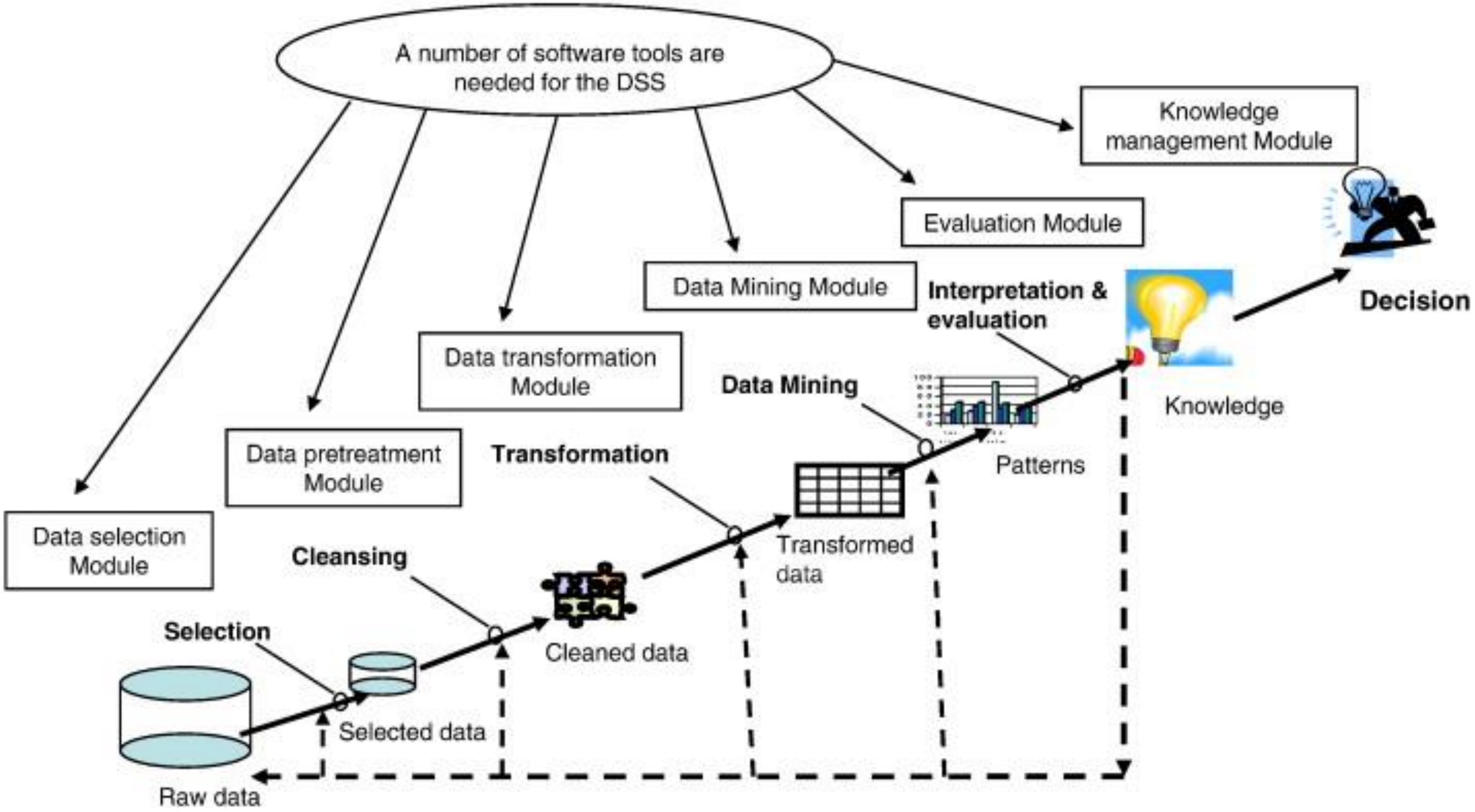
m : weight to prior (number of "virtual" examples, $m\geq 1$)

Zero conditional probability

- Example: P(outlook=overcast|no)=0 in the play-tennis dataset
  - Adding m “virtual” examples (m: up to 1% of #training example)
    - In this dataset, # of training examples for the “no” class is 5.
    - We can only add m**=1** “virtual” example in our m-esitmate remedy.
  - The “outlook” feature can takes only 3 values. So $p=1/3$.
  - Re-estimate P(outlook|no) with the m-estimate

$$\mathrm{P(overcast|no)} = \frac{0+1*\left(\frac{1}{3}\right)}{5+1} = \frac{1}{18}$$

$$\mathrm{P(sunny|no)} = \frac{3+1*\left(\frac{1}{3}\right)}{5+1} = \frac{5}{9} \qquad \mathrm{P(rain|no)} = \frac{2+1*\left(\frac{1}{3}\right)}{5+1} = \frac{7}{18}$$



Ayed, B. M., Ltifi, H., Kolski,C. & Alimi, A. (2010) A usercentered approach for the design & implementation of KDD-based DSS: A case study in the healthcare domain. Decision Support Systems, 50, 64- 78.

Evaluation

- The 2x2 Table
- The Gold Standard
- Sensitivity and Specifity
- Tests with High Sensitivity
- Tests with High Specifity
- ROC Curve
- Area Under The ROC Curve (AUC)
- Predictive Values (Positive and Negative)
- Prevalance

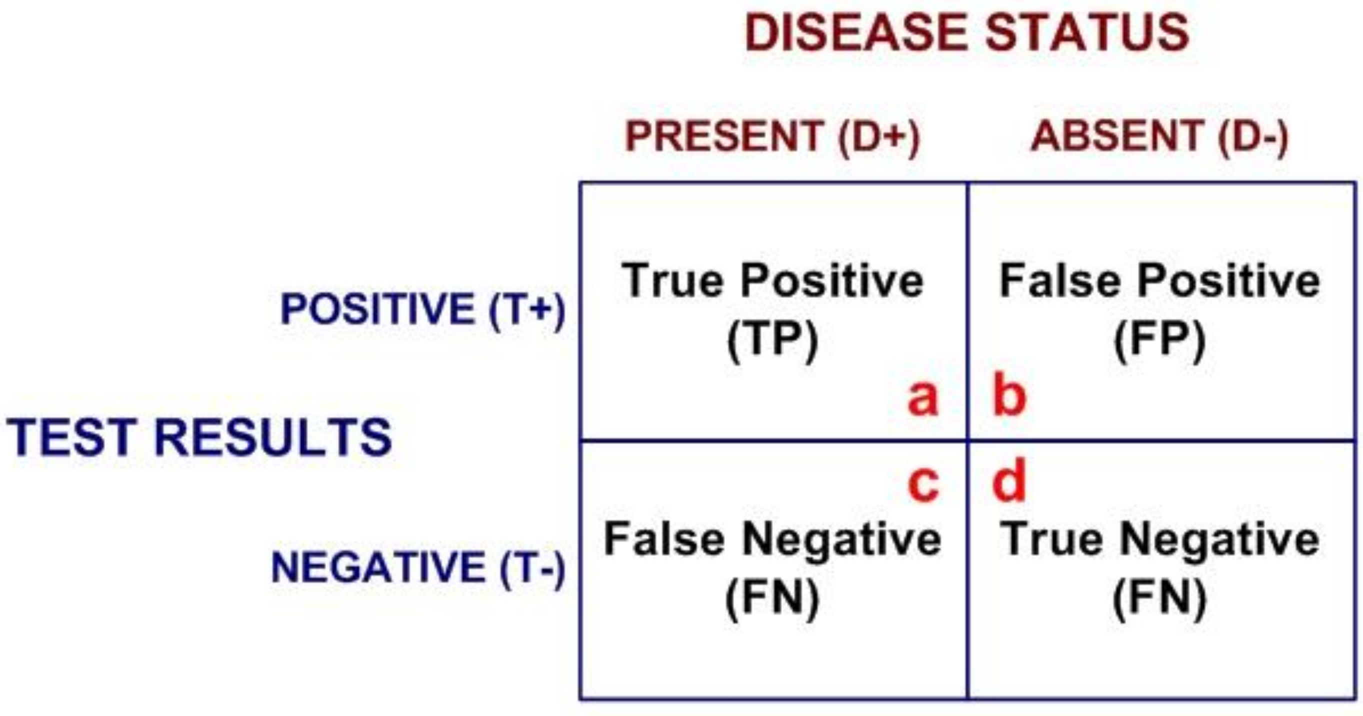
CLINICAL DIAGNOSIS AND CLINICAL TESTING

- Diagnosis: the process of discovering a patient’s underlying disease status by:
  - ascertaining the patient’s history, signs and symptoms, choosing appropriate tests, interpreting the results, and making correct conclusions.
  - highly complicated, not well understood process.
- Testing: the application of clinical test information to infer disease status:

Clinical test information refers to any piece of information not just laboratory or diagnostic tests!!

THE 2 X 2 TABLE

Relationship between Diagnostic Test Result and Disease Status



		DISEASE STATUS	
		PRESENT (D+)	ABSENT (D-)
TEST RESULTS	POSITIVE (T+)	True Positive (TP) a	b False Positive (FP)
	NEGATIVE (T-)	c False Negative (FN)	d True Negative (FN)

THE GOLD STANDARD (or REFERENT STANDARD)

- Definition: The accepted standard for determining the true disease status.
- Want to know the disease status with certainty but this is frequently not possible because:
  - Gold standard test is difficult, expensive, risky, unethical, or simply not possible
    - e.g., DVT requires leg venogram (difficult, expensive)
    - e.g., vCJD requires autopsy (not possible)
  - Frequently resort to using an imperfect proxy as a referent standard
    - e.g., Ultrasound and/or 3-month follow-up in place of venogram to confirm presence/absence of DVT.

A bit «TERMINOLOGY»…

- Machine learning;

$$\text{True positive rate} = \frac{TP}{TP + FN} \quad \text{true negative rate} = \frac{TN}{TN + FP}$$

$$\text{(total) accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{N}$$

$$\text{positive predictive value} = \frac{TP}{TP + FP} \quad \text{negative predictive value} = \frac{TN}{TN + FN}$$

- Medical informatics & Clinical research;

sensitivity = true positive rate   specificity = true negative rate

A bit «TERMINOLOGY»…

- Information retrieval;

$$\text{precision} = \text{positive predictive value} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = p$$

$$\text{F measure} = \frac{2pr}{p+r}$$

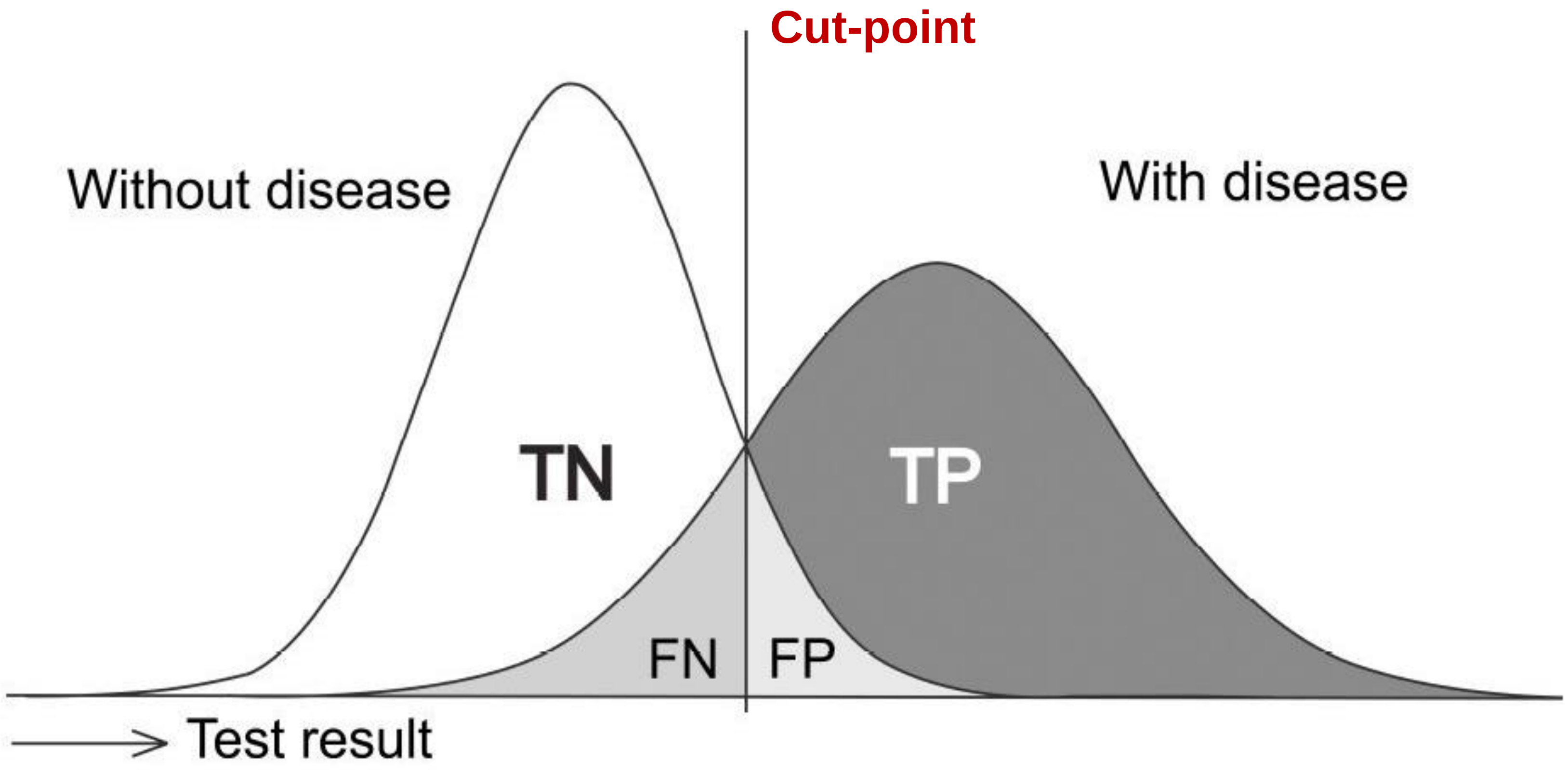
$$\text{recall} = \text{true postive rate} = \text{sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = r$$

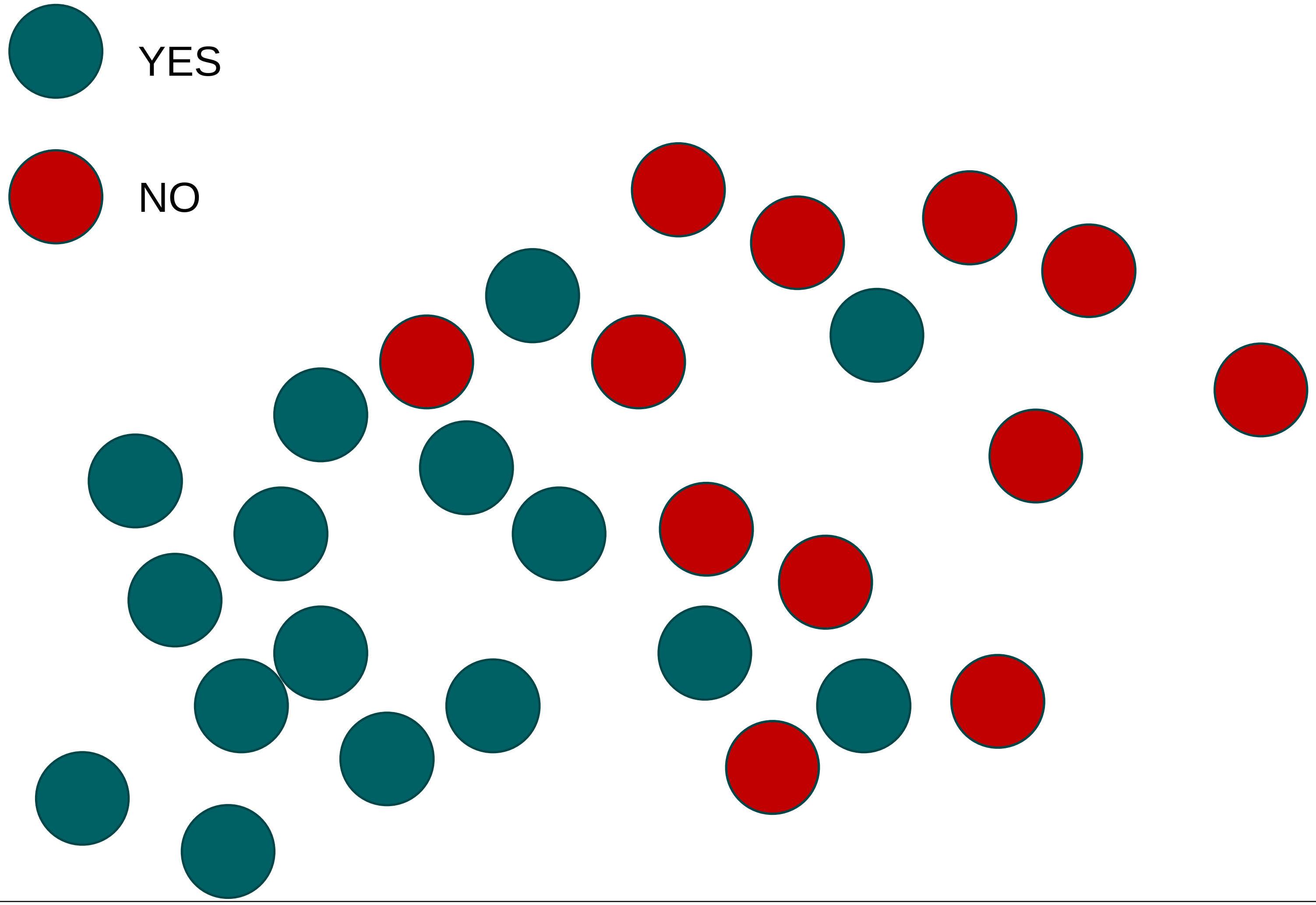
SENSITIVITY (Se) & SPECIFICITY (Sp)

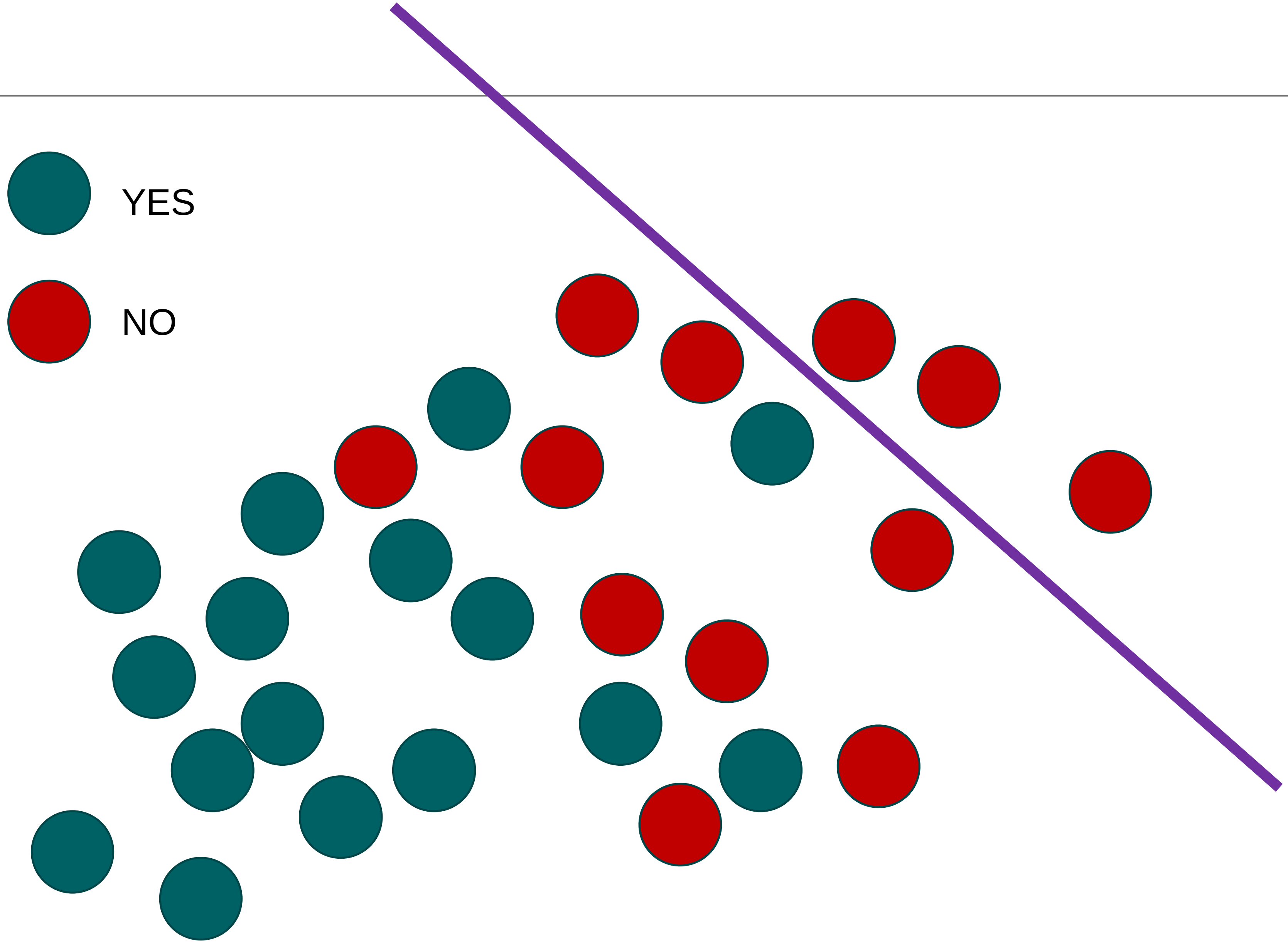
- Interpretation of diagnostic tests is concerned with comparing the relative frequencies and “costs” of the incorrect results (FNs and FPs) versus the correct results (TPs and TNs).
- Degree of overlap is a measure of the test’s effectiveness or discriminating ability which is quantified by Se and Sp.

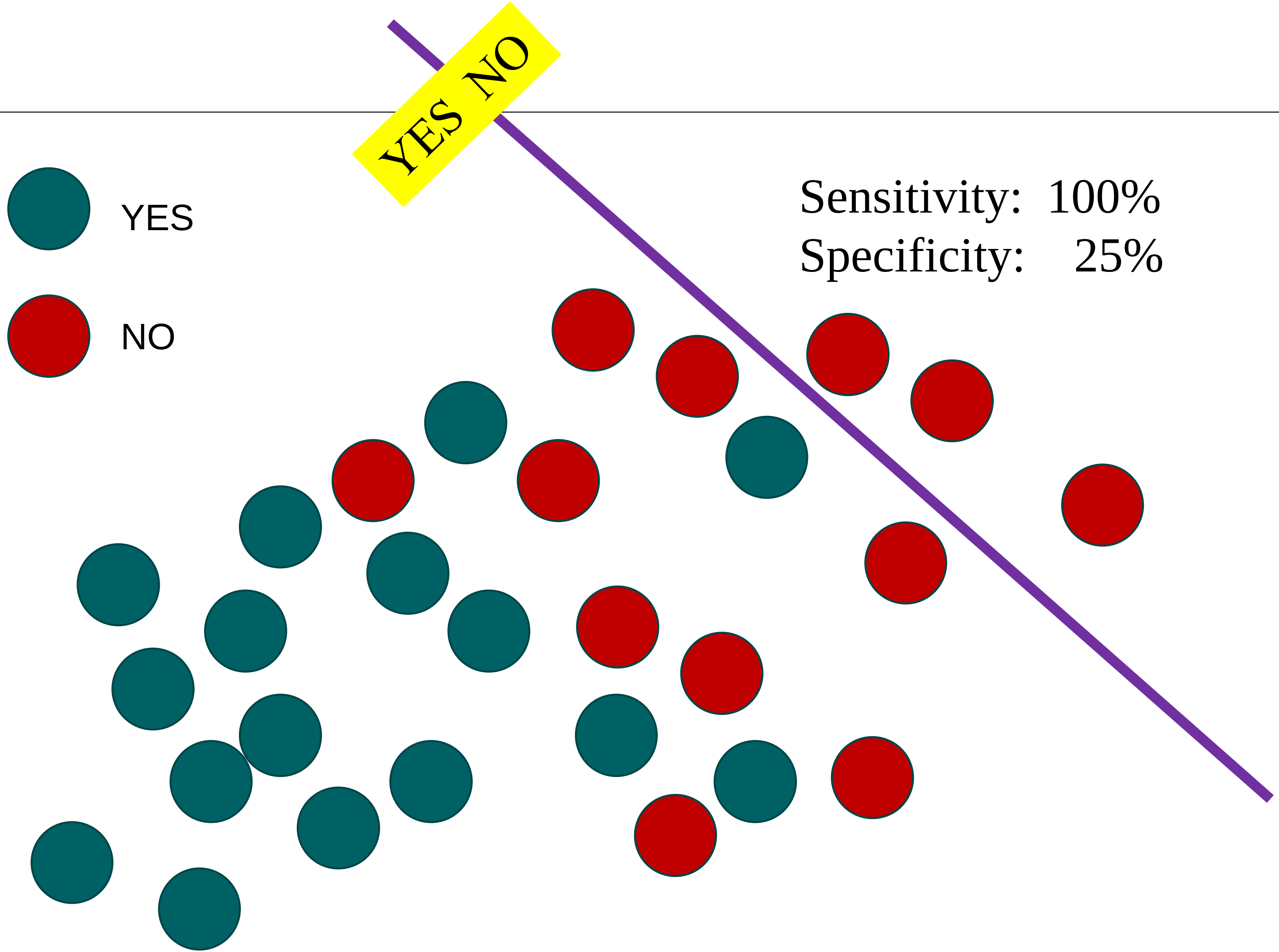
SENSITIVITY (Se) & SPECIFICITY (Sp)

Results for a Typical Diagnostic Test Illustrating Overlap Between Disease (D+) and Non-disease (D-) Populations



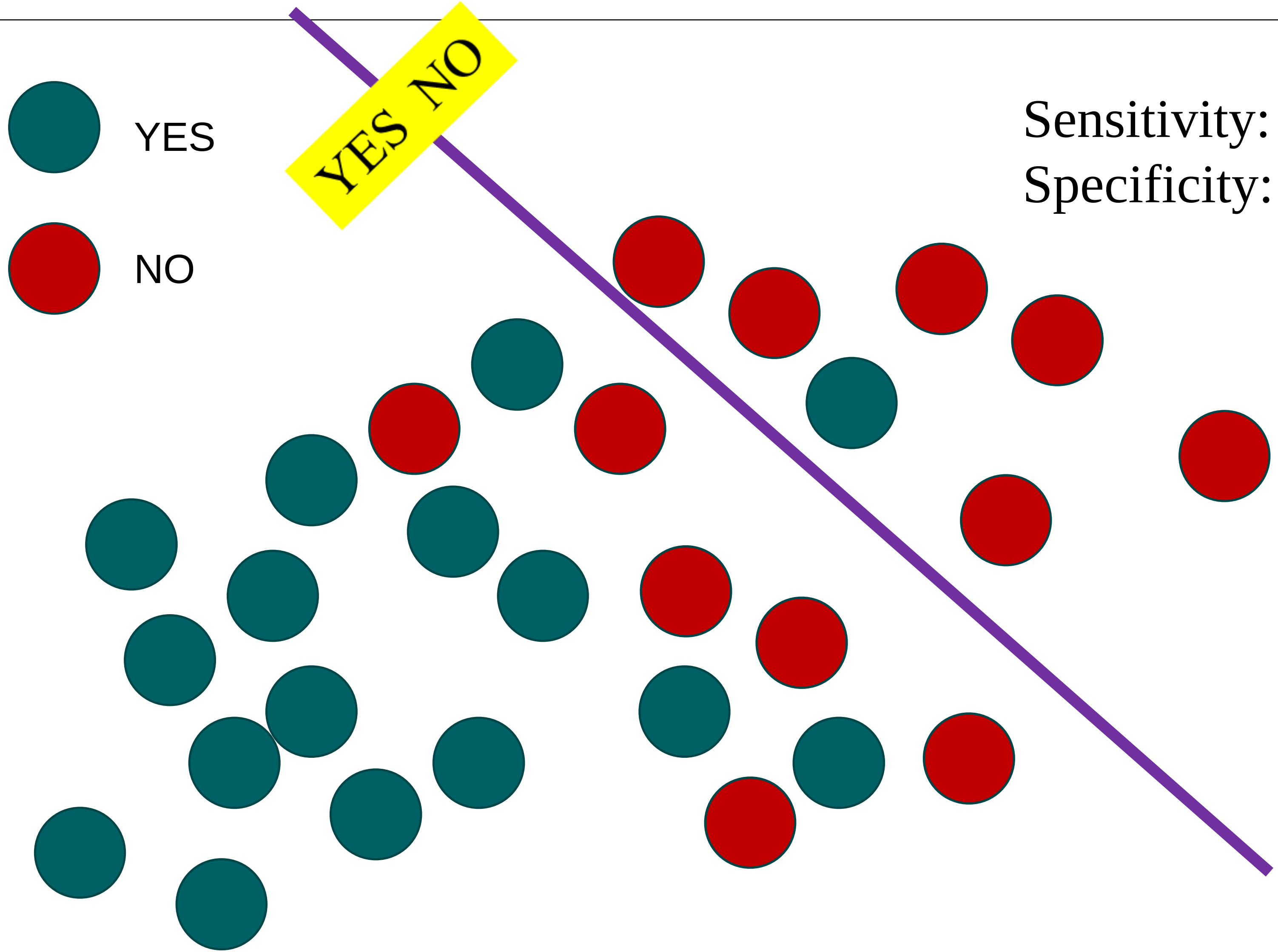

YES
NO


YES
NO



Sensitivity: 100%

Specificity: 25%

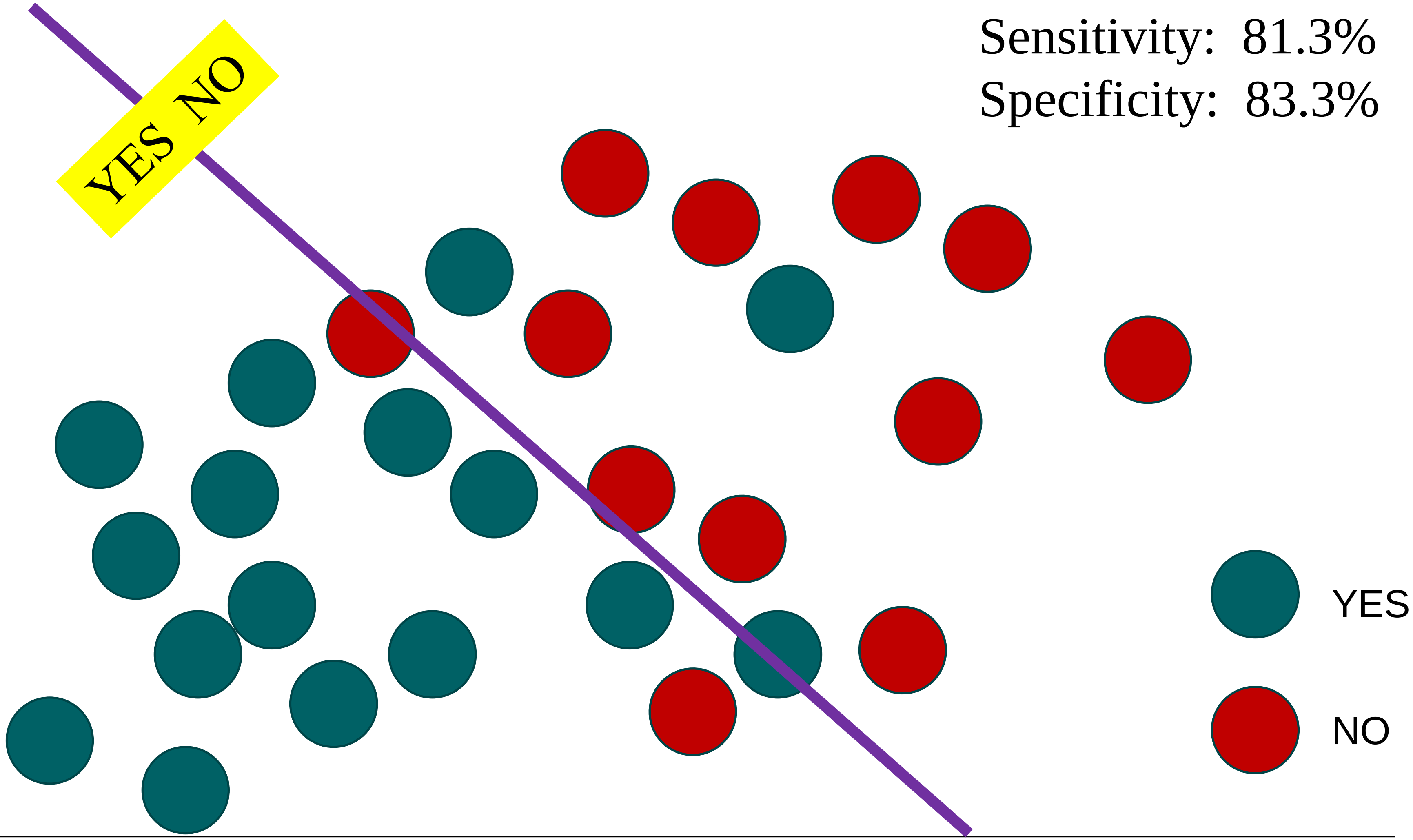


Sensitivity: 93.8%

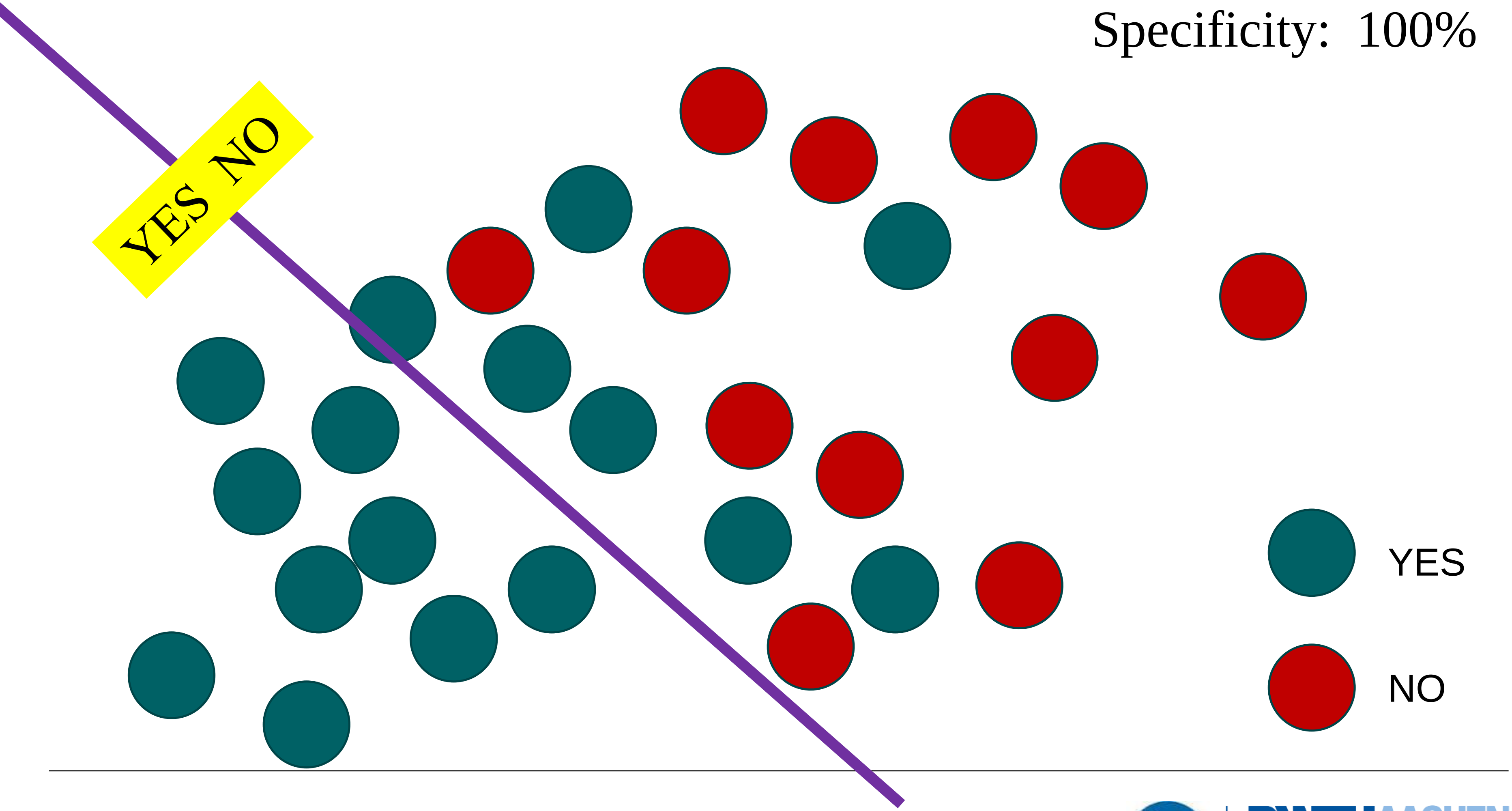
Specificity: 50%

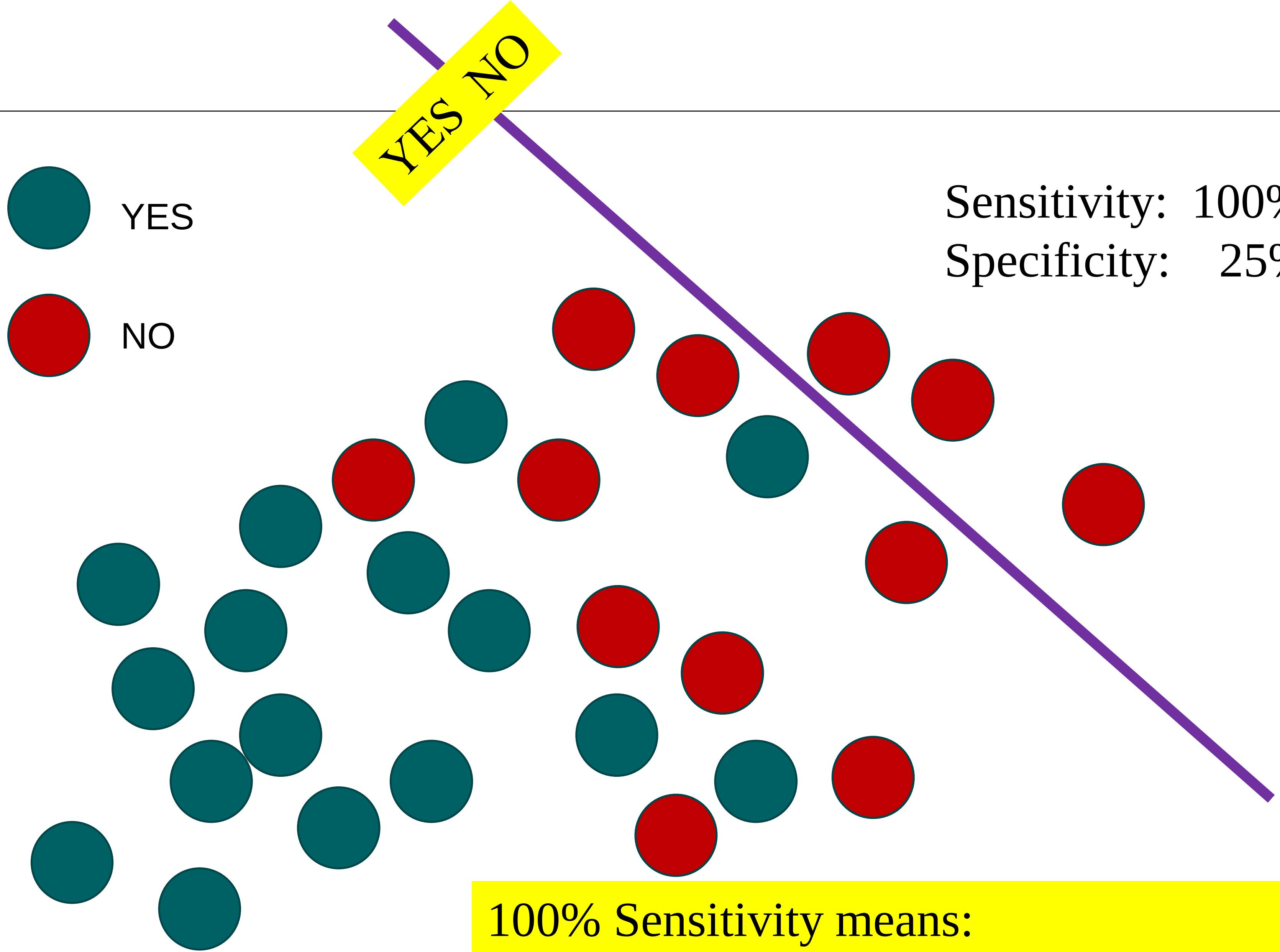
Sensitivity: 81.3%

Specificity: 83.3%



Sensitivity: 56.3%
Specificity: 100%



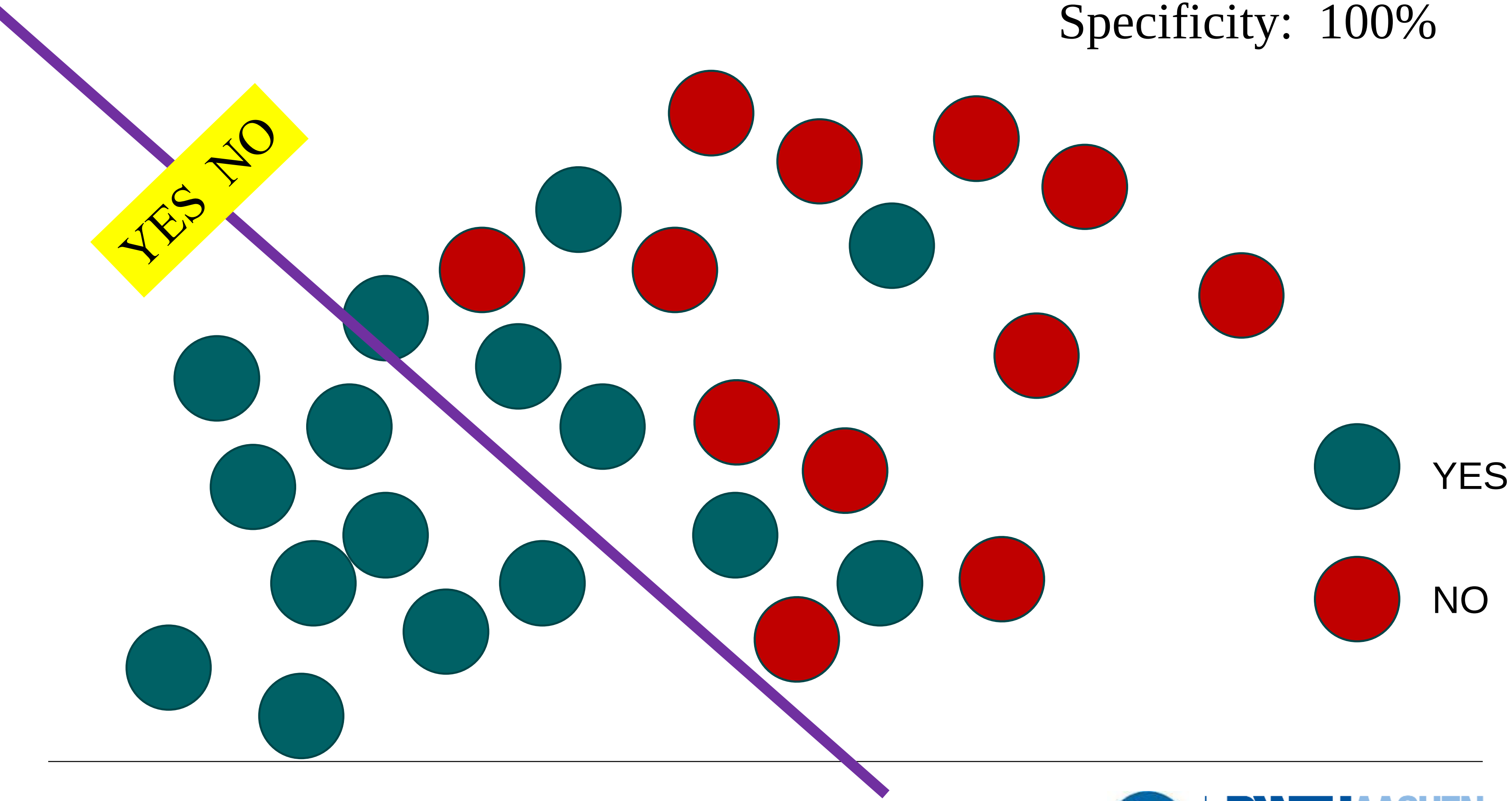


Sensitivity: 100%
Specificity: 25%

100% Sensitivity means:
detects ***all*** cancer cases (or whatever)
but possibly with many false positives

100% Specificity means:
misses some cancer cases (or whatever)
but no false positives

Sensitivity: 56.3%
Specificity: 100%



SENSITIVITY (Se)

- Definition: the proportion of individuals with disease that have a positive test result, or

$$Se = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} = \frac{TP}{TP + FN} = \frac{a}{a + c}$$

- Se; conditional probability of being test positive given that disease is present

$$Se = P(T+ \mid D+)$$

- calculated solely from diseased individuals
- also referred to as the true-positive rate
- a.k.a = “the probability of calling a case a case”

CALCULATION OF SENSITIVITY (Se) & SPECIFICITY (Sp)

Se and Sp are calculated from the left and right columns, respectively

		DISEASE STATUS	
		PRESENT (D+)	ABSENT (D-)
TEST RESULTS	POSITIVE (T+)	True Positive (TP) a	b False Positive (FP)
	NEGATIVE (T-)	c False Negative (FN)	d True Negative (TN)

Se = TP/TP+FN **Sp = TN/TN+FP**

Se = a / a + c **Sp = d / d + b**

SPECIFICITY (Sp)

- Definition: the proportion of individuals without disease that have a negative test result, or

$$Sp = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} = \frac{TN}{TN + FP} = \frac{d}{d + b}$$

- conditional probability of being test negative given that disease is absent

$$Sp = P(T\text{-}|D\text{-})$$

- calculated solely from non-diseased individuals
- also referred to as the true-negative rate
- a.k.a = “the probability of calling a control a control”

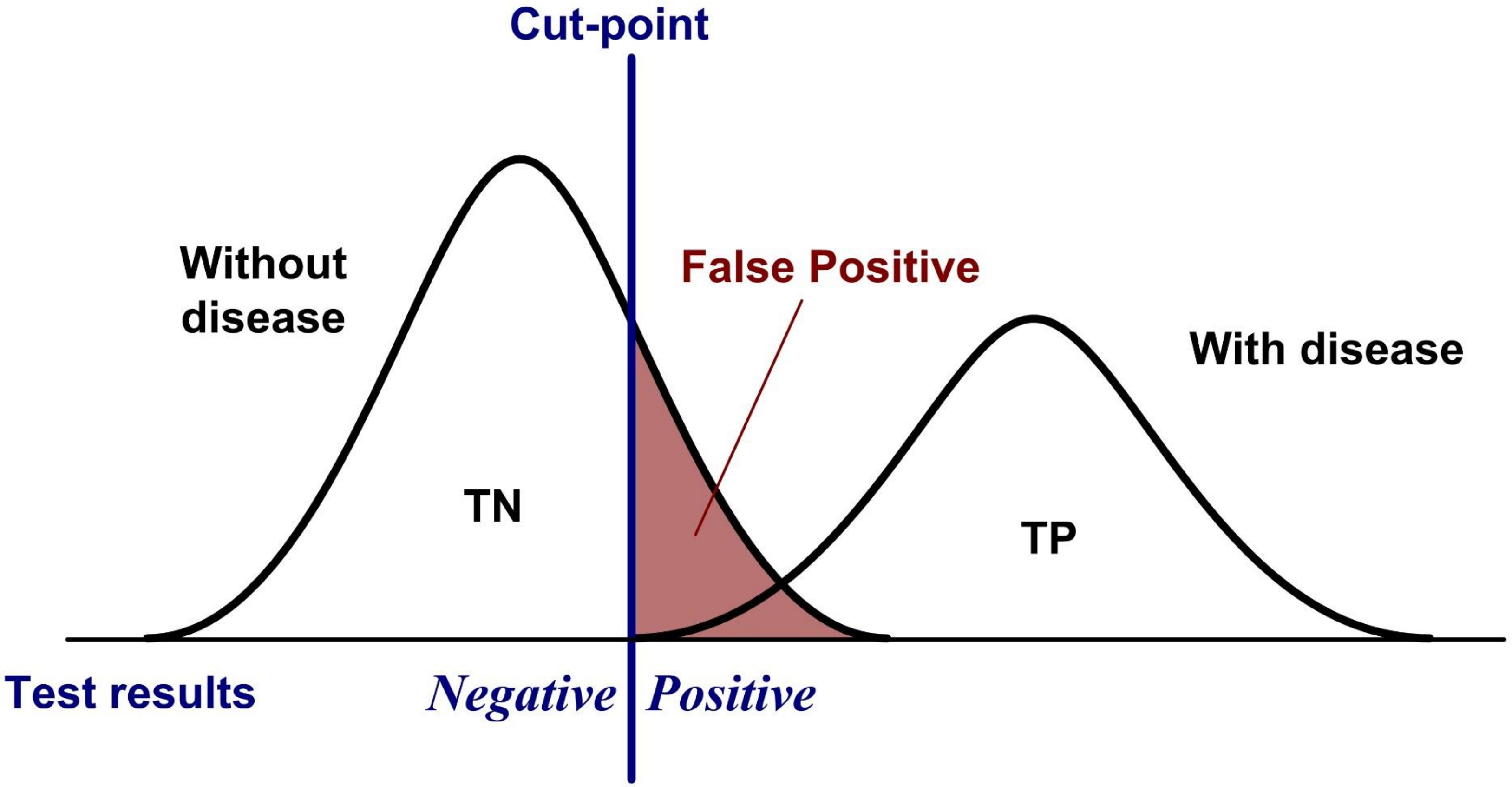
TESTS WITH HIGH SENSITIVITY

- Perfectly sensitive test (Se = 100%), all diseased patients test positive (no FN's), therefore
  - All test negative patients are disease free (TNs)
  - Typical tradeoff is a decreased Sp (many FPs)
- Highly sensitive tests are used to rule-out disease
  - if the test is negative you can be confident that disease is absent (FN results are rare!)

SnNOUT = if a test has a sufficiently high ***Sen****sitivity, a* ***N****egative result rules* ***OUT*** *disease*

TESTS WITH HIGH SENSITIVITY

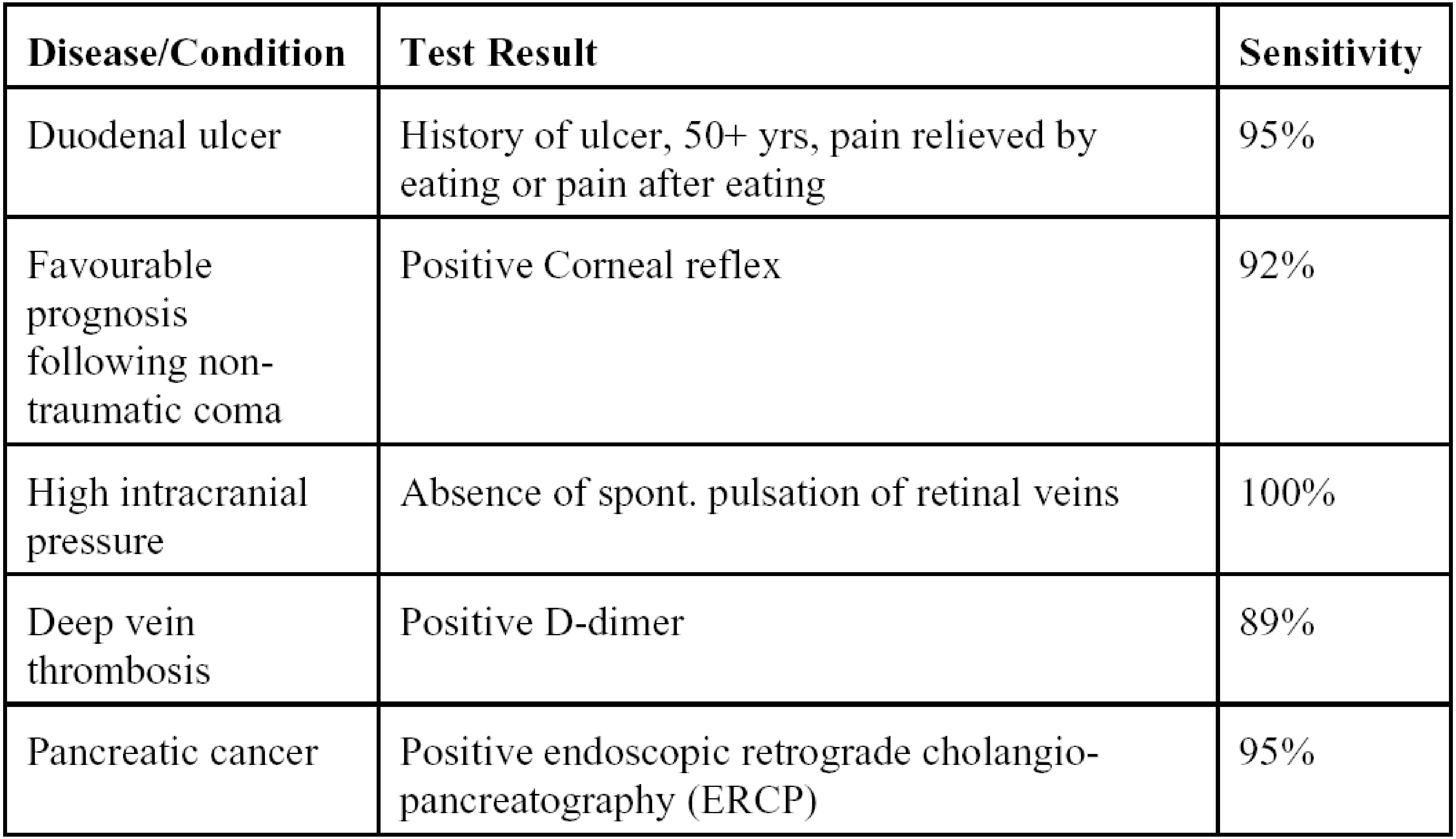
Example of a Perfectly Sensitive Test (no FNs)



CLINICAL APPLICATIONS OF TESTS WITH HIGH Se

- Early stages of a diagnostic work-up.
  - large number of potential diseases are being considered.
  - a negative result indicates a particular disease can be dropped (i.e., ruled out).
- Important penalty for missing a disease.
  - dangerous but treatable conditions e.g., DVT, TB, syphilis
  - don’t want to miss cases, hence avoid false negative results
- Screening tests.
  - the probability of disease is relatively low (i.e., low prevalence)
  - want to find as many asymptomatic cases as possible (increased ‘yield’ of screening)

EXAMPLES OF TESTS WITH HIGH Se



Disease/Condition	Test Result	Sensitivity
Duodenal ulcer	History of ulcer, 50+ yrs, pain relieved by eating or pain after eating	95%
Favourable prognosis following non-traumatic coma	Positive Corneal reflex	92%
High intracranial pressure	Absence of spont. pulsation of retinal veins	100%
Deep vein thrombosis	Positive D-dimer	89%
Pancreatic cancer	Positive endoscopic retrograde cholangio-pancreatography (ERCP)	95%

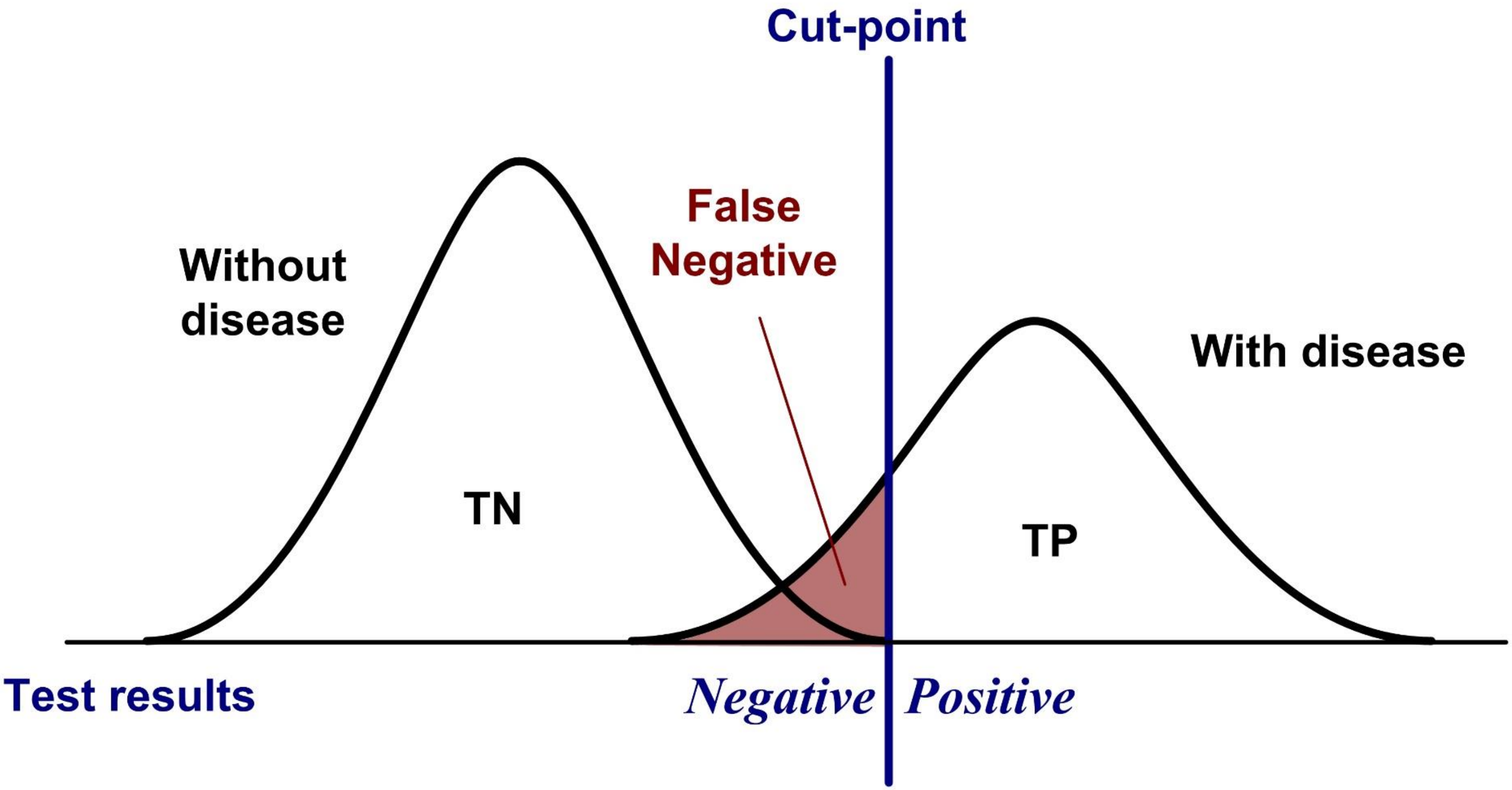
TESTS WITH HIGH SPECIFICITY

- Perfectly specific test (Sp = 100%), all non-diseased patients test negative (no FP's), therefore
  - All test positive patients have disease (TPs)
  - Typical tradeoff is large number of FNs
- Highly specific tests are used to rule-in disease
  - if the test is positive you can be confident that disease is present (FPs are rare).

*SpPin = if a test has a sufficiently high **Sp**ecificity, a **P**ositive result rules **in** disease.*

EXAMPLE OF A PERFECTLY SPECIFIC TEST

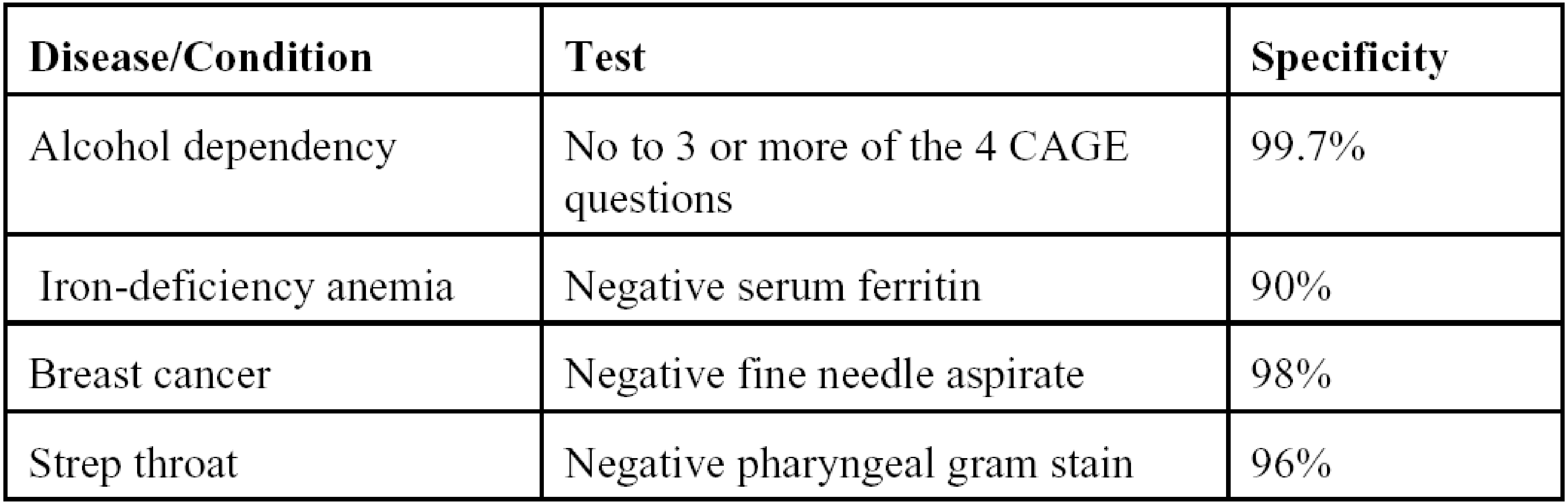
Example of a Perfectly Specific Test (no FPs)



CLINICAL APPLICATIONS OF TESTS WITH HIGH Sp

- To rule-in a diagnosis suggested by other tests
  - specific tests are therefore used at the end of a work-up to rule-in a final diagnosis e.g., biopsy, culture.
- False positive tests results can harm patient
  - want to be absolutely sure that disease is present.
  - example, the confirmation of HIV positive status or the confirmation of cancer prior to chemotherapy.

EXAMPLES OF TESTS WITH HIGH Sp



Disease/Condition	Test	Specificity
Alcohol dependency	No to 3 or more of the 4 CAGE questions	99.7%
Iron-deficiency anemia	Negative serum ferritin	90%
Breast cancer	Negative fine needle aspirate	98%
Strep throat	Negative pharyngeal gram stain	96%

TRADE OFF BETWEEN Se AND Sp

- Obviously we’d like tests with both high Se and Sp (> 95%), but this is rarely possible
- An inherent trade-off exists between Se and Sp (if you increase one the other must decrease)
- Whenever clinical data take on a range of values the location of the cut-point is arbitrary
  - Location should depend on the purpose of the test
  - Methods exist to calculate the best cut-point based on the frequency and relative “costs” of the FN and FP results
  - Trade-off between Se and Sp is demonstrated on ROC curve

Sensitivity and Specificity in Model Building

When you are training your model, what should be the trade off between sensitivity and specificity .

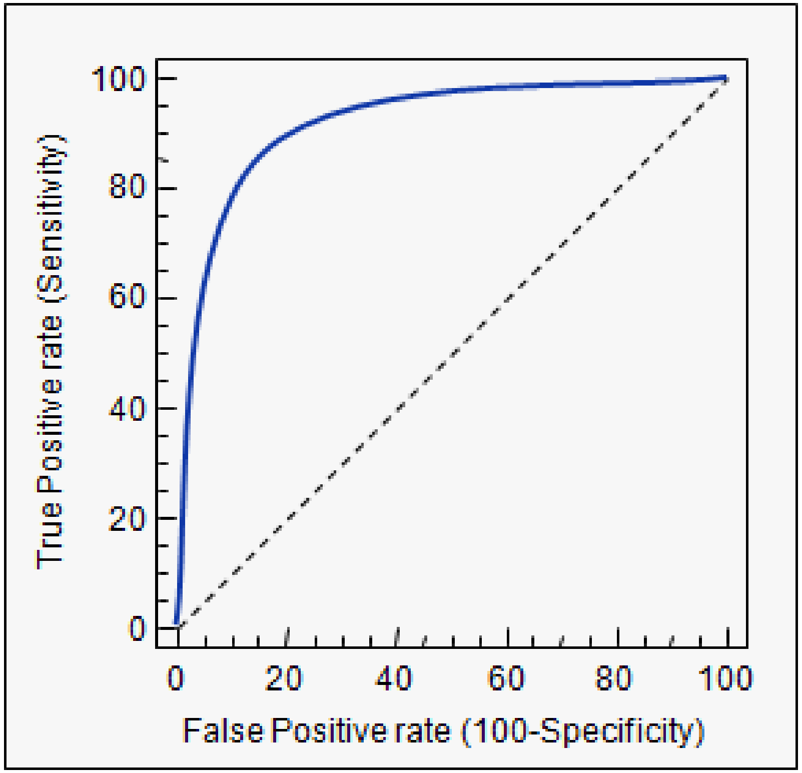
Discuss the following cases:

- You used TCGA data repository (includes various genomic data and patient records) to discover common underlying genetic mutations for diabetes and cardiovascular diseases. You found some common gene expressions. Based on your initial findings, molecular biologist will conduct further vet lab experiments

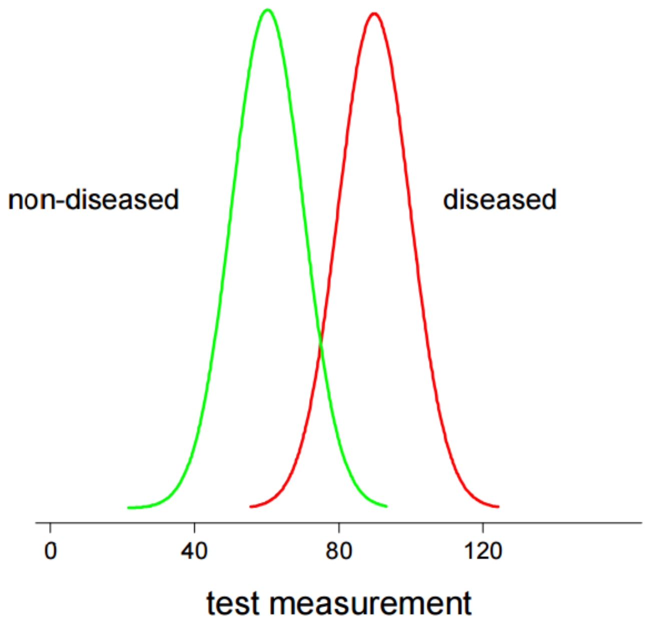
- You are using data from ambient sensing and wearable devices data to detect anomalies in elderly daily behavior, such as falls. In case you detect a fall, emergency services will run into elderly patients home and rescue them.

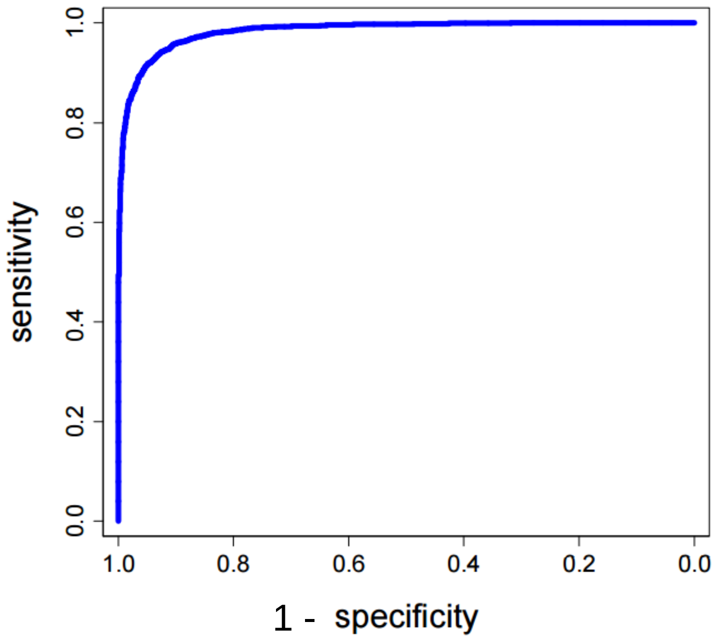
ROC CURVE

- The diagnostic performance of a test, or the accuray of a test to discriminate diseased cases from normal cases is evaluated using Receiver Operating Characteristic (ROC) curve analysis.
- In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter.
- Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.



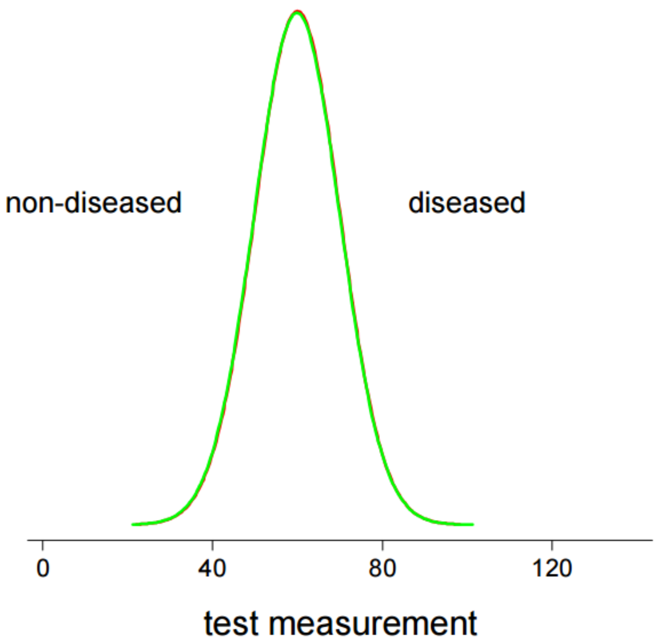
EXAMPLES OF ROC CURVES

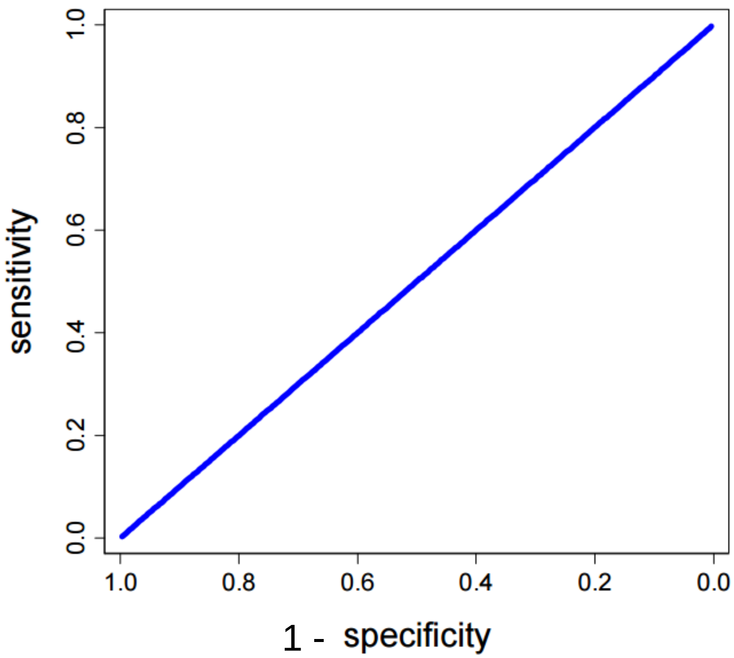




- The position of the ROC curve depends on the degree of overlap of the distributions of the test measurement in diseased and non-diseased.
- Where a test clearly discriminates between diseased and non-diseased, the ROC curve will indicate that high sensitivity is achieved with a high specificity, that is the curve approaches the upper left hand corner of the graph where sensitivity is 1 and specificity is 1.

EXAMPLES OF ROC CURVES

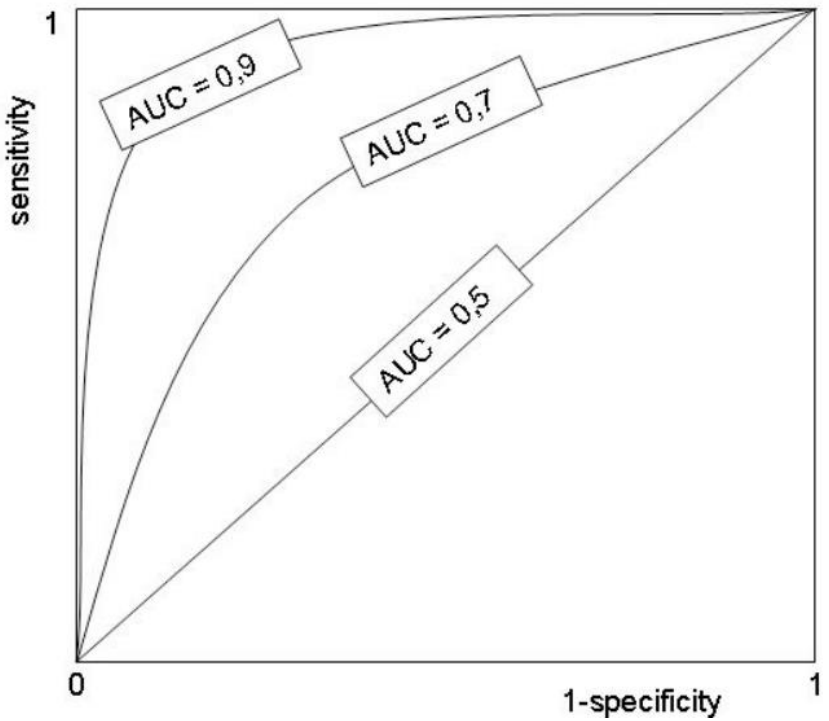




- If the distributions of test results in diseased and non-diseased coincide, the test would be completely uninformative and its ROC curve would be the upward diagonal of the square.

AREA UNDER THE ROC CURVE (AUC)

- The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal).



AUC	diagnostic accuracy
0.9 – 1.0	excellent
0.8 - 0.9	very good
0.7 - 0.8	good
0.6 - 0.7	sufficient
0.5 - 0.6	bad
< 0.5	test not useful

PREDICTIVE VALUES

- In terms of conditional probabilities Se and Sp are defined as:

  $Se = P(T+|D+)$ $Sp = P(T-|D-)$

- **Problem:** can only be calculated if the true disease status is known!
- But the clinician is using a test precisely because the disease status is unknown!
- So, clinician actually wants the conditional probability of disease given the test result, OR $P(D+|T+)$ and $P(D-|T-)$

PREDICTIVE VALUE POSITIVE (PVP)

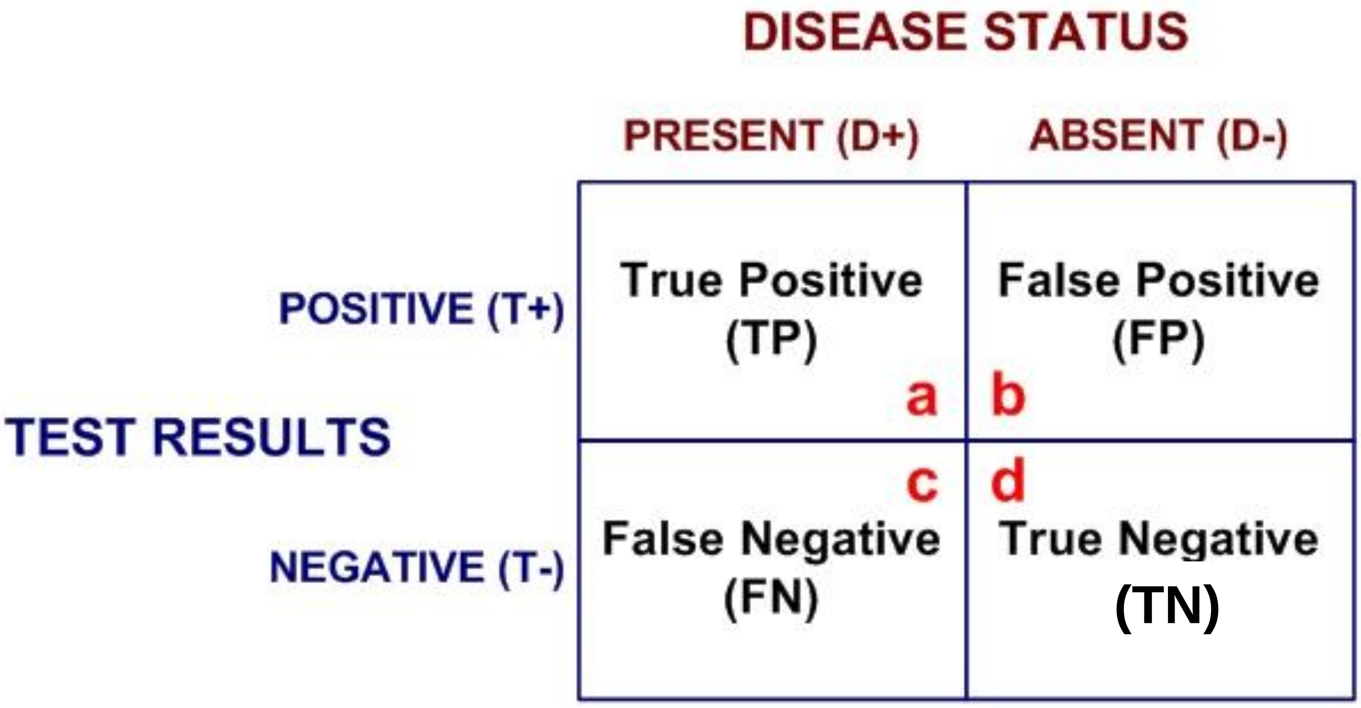
- Definition: The probability of disease in a patient with a positive (abnormal) test.

$$\text{PVP} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{a}{a + b}$$

- calculated solely from test positive individuals (top row of 2 x 2 table)
- conditional probability of being diseased given the test was positive, or PVP = P(D+|T+)
- Sp and PVP are linked and provide information on the FP rate. A highly specific test helps to rule-in disease because PVP is maximized.

CALCULATION OF PREDICTIVE VALUES

PVP and PVN are calculated from the top and bottom rows, respectively



		DISEASE STATUS	
		PRESENT (D+)	ABSENT (D-)
TEST RESULTS	POSITIVE (T+)	True Positive (TP) a	b False Positive (FP)
	NEGATIVE (T-)	c False Negative (FN)	d True Negative (TN)

PVP = TP/TP+FP

PVP = a / a + b

PVN = TN/TN+FN

PVN = d / d + c

PREDICTIVE VALUE NEGATIVE (PVN)

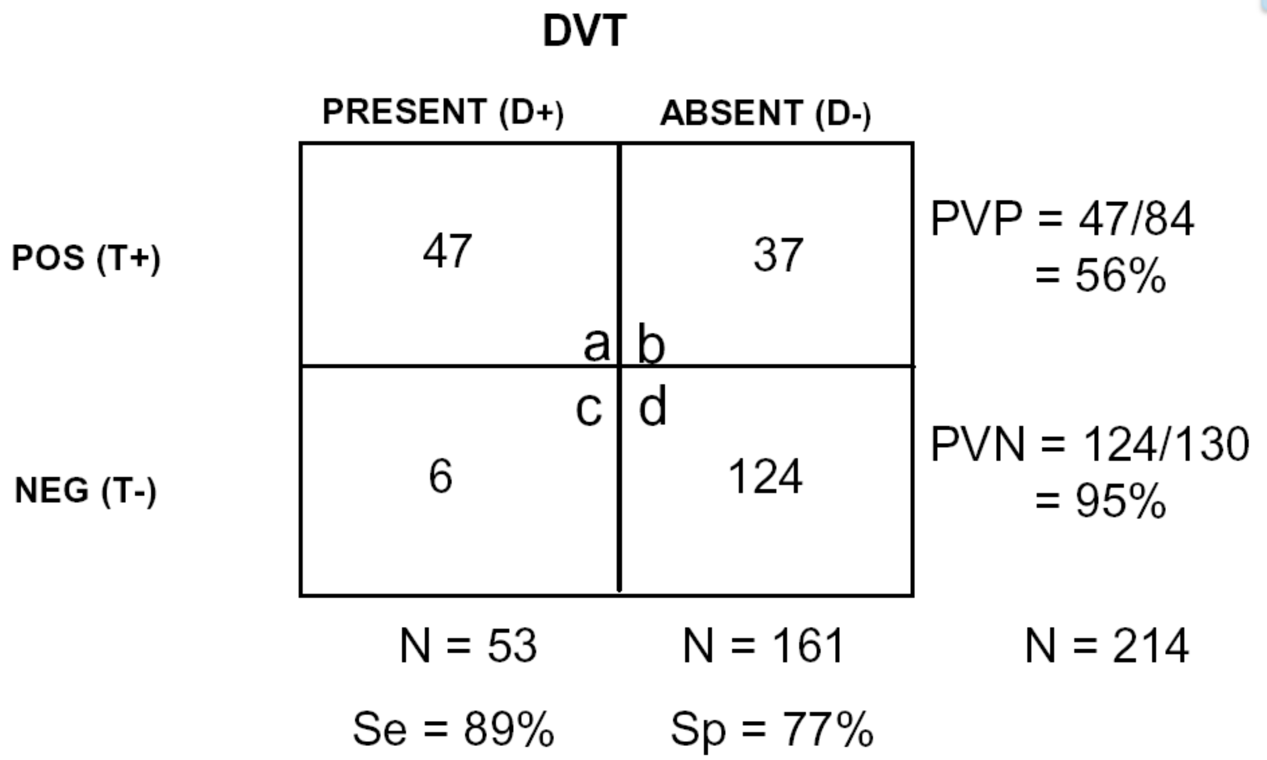
- Definition: The probability of not having disease when the test result is negative (normal).

$$\text{PVN} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}} = \frac{\text{TN}}{\text{TN} + \text{FN}} = \frac{\text{d}}{\text{d} + \text{c}}$$

- calculated solely from test negative individuals (bottom row of 2 x 2 table).
- conditional probability of not being diseased given the test was negative or PVN = P(D-|T-)
- Se and PVN are linked and provide information on the FN rate. A highly specific test helps to rule-out disease because PVN is maximized.

COMPARISON OF PVP and PVN

The PVP and PVN of D-dimer whole blood assay (SimpliRED assay) for DVT (Ref: Wells PS, Circulation, 1995). Prevalence = 25%



DVT

	PRESENT (D+)	ABSENT (D-)	
POS (T+)	47 a	37 b	PVP = 47/84 = 56%
NEG (T-)	c 6	d 124	PVN = 124/130 = 95%
	N = 53	N = 161	N = 214
	Se = 89%	Sp = 77%	

PREVALENCE

- the proportion of the total population tested that have disease, or

$$P = \frac{\text{Total Number of Diseased}}{\text{Total Population (N)}} = \frac{TP + FN}{TP+FN+FP+TN} = \frac{a + c}{a + b + c + d}$$

- Equivalent names: prior probability, the likelihood of disease, prior belief, prior odds, pre-test probability, and pre-test odds.

IMPORTANCE OF PREVALENCE

The PVP and PVN of D-dimer whole blood assay (SimpliRED assay) for DVT (Ref: Wells PS, Circulation, 1995). Prevalence = 5%

DVT

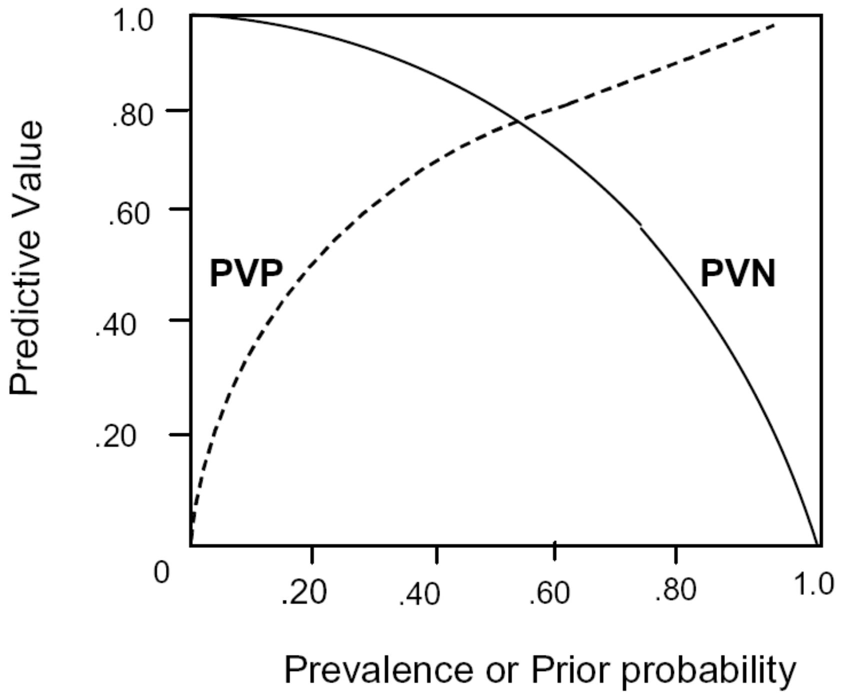
	PRESENT (D+)	**ABSENT (D-)**	
POS (T+)	10 a	47 b	PVP = 10/57 = 18%
NEG (T-)	c 1	d 156	PVN = 156/157 = 99.4%
	N = 11	N = 203	N = 214
	Se = 89%	Sp = 77%	

IMPORTANCE OF PREVALENCE

- Has a dramatic influence on predictive values
- Prevalence can vary widely from hospital to hospital, clinic to clinic, or patient to patient
- The same test (meaning the same Se and Sp) when applied under different scenarios (meaning different prevalence's) can give very different results (meaning different PVP and PVN!)
- Prior probability represents what the clinician believes (prior belief or clinical suspicion)
- set by considering the practice environment, patients history, physical examination findings, experience and judgment etc.

IMPORTANCE OF PREVALENCE

The PVP and PVN as a Function of Prevalence for a Typical Diagnostic Test



As prevalence falls, positive predictive value must fall along with it, and negative predictive value must rise. Conversely, as prevalence increases, positive predictive value will increase and negative predictive value will fall.

REFERENCES

- Barber, David. Bayesian reasoning and machine learning. Cambridge University Press, 2012.
- Ana-Maria Šimundić. **Measures of diagnostic accuracy: basic definitions.** http://www.ifcc.org/ifccfiles/docs/190404200805.pdf
- Petra Macaskill, Constantine Gatsonis, Jonathan Deeks, Roger Harbord, Yemisi Takwoingi. **Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy** Chapter:10 Analysing and Presenting Results. http://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter%2010%20-%20Version%201.0.pdf
- Milos Jenicek. **A Primer on Clinical Experience in Medicine, Reasoning, Decision.** CRC Press. 2013.
- **ROC curve analysis in MedCalc.** https://www.medcalc.org/manual/roc-curves.php