

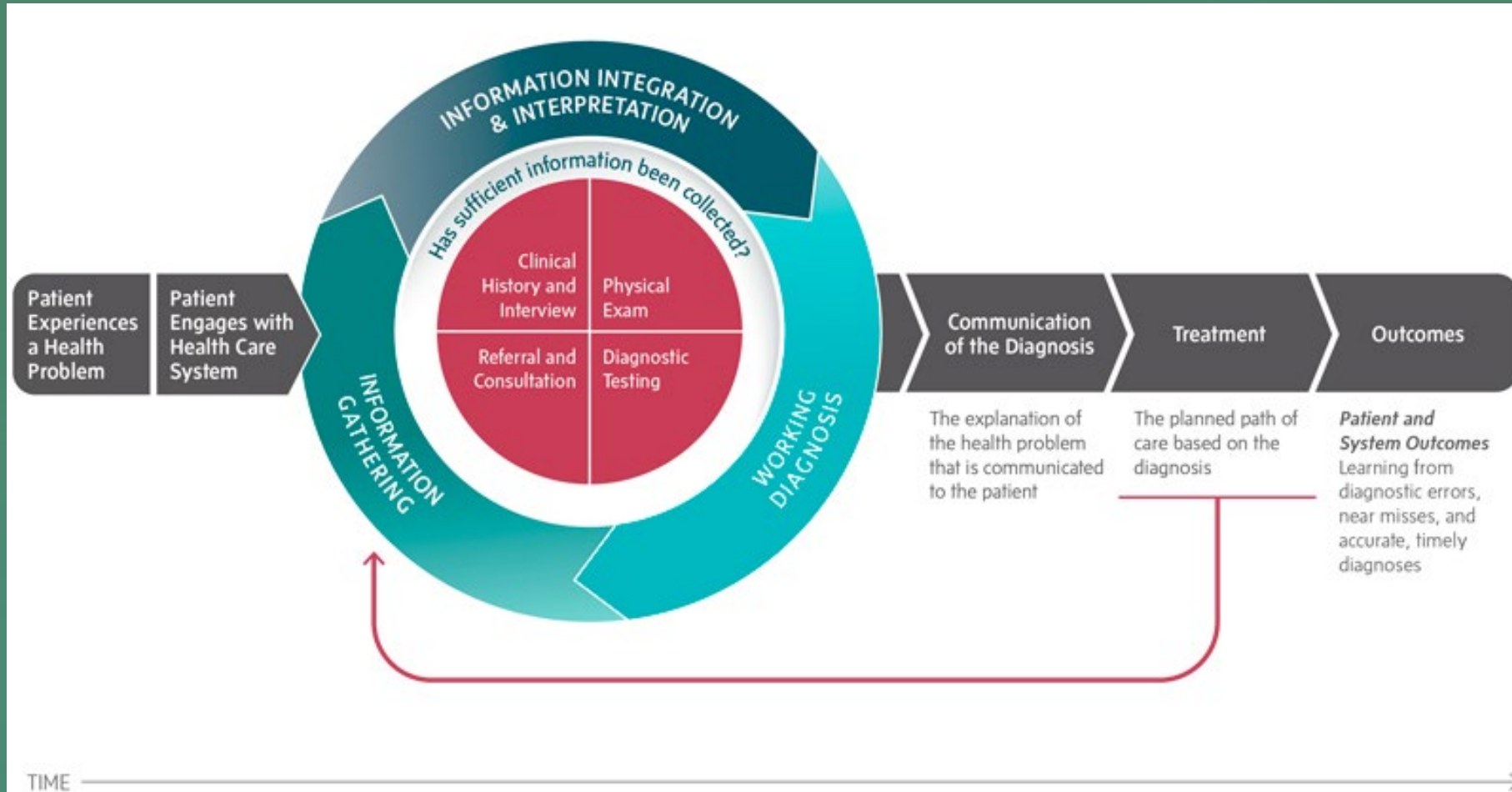
# STUDY DESIGNS

## IN BIOMEDICAL RESEARCH




# BIOMARKER RESEARCH

# The Complete Healthcare Process



The National Academies of  
SCIENCES • ENGINEERING • MEDICINE

SOURCE: National Academies of Sciences, Engineering, and Medicine. 2015.  
*Improving Diagnosis in Health Care*. Washington, DC: The National Academies Press.



An important part of the healthcare process, an crucial station, is Disease Diagnosis. It is part of Translational Research – the intersection between Basic Science and Clinical Science (Medicine). Some of us in Statistics and Biostatistics called a section of this part “**Biomarker Research**”.

# Diagnostic Biomarkers

**Definition:** A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions

**Types:** Molecular, histologic, radiographic, or physiologic characteristics

**Examples:**

1. Prostate specific antigen (PSA) for prostate cancer (Molecular)
2. Estrogen receptor (ER), Progesterone receptor (PR), and HER-2 for breast cancer (Molecular)
3. Gleason score for prostate cancer (histologic)
4. Mammogram score (BI-RADS) for breast cancer (radiographic)
5. Blood pressure for high blood pressure (physiologic characteristics)
6. BMI for obesity (physiological characteristics)

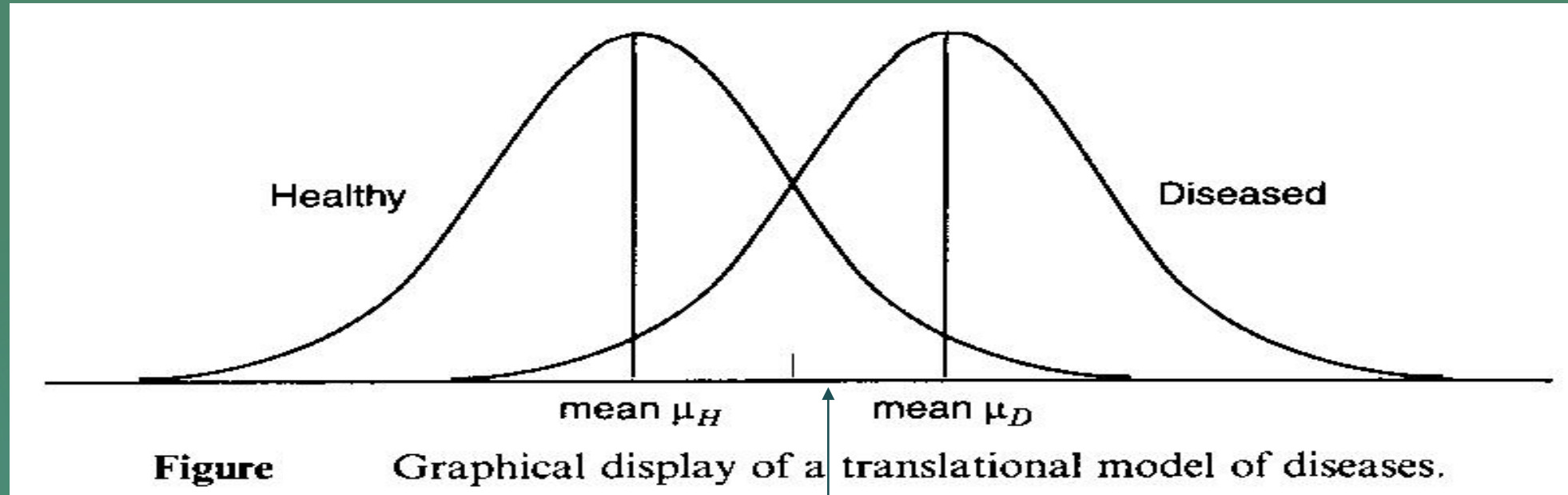
**Common data features:** Either **CONTINUOUS** or **ORDINAL**

# ROC CURVE

Diagnostic tests have been presented as always having dichotomous outcomes. In some cases, the result of the test may be binary, but in many cases it is based on the dichotomization of a continuous biomarker – some factor correlated to the absence or presence of the disease.

To deal with a continuous biomarker, we need a well-known graph called the Receiver Operating Characteristic curve or “ROC curve”.

# A SIMPLE MODEL & DICHOTOMIZATION



Biomarker Y is normally distributed with the same variance, but different means; no matter where you “cut”, both errors result! Also, specificity & sensitivity are functions of the “cutpoint” y.

# ASSUMPTION FOR SIMPLIFICATION

- ▶ In the case of many diseases, the **larger values of the biomarker Y are associated with the diseased population**; e.g. blood glucose for diabetes, PSA for prostate cancer),
- ▶ For many others, the **smaller values of the biomarker Y are associated with the diseased population**; static admittance for Otitis Media, TSH for hyperthyroidism).
- ▶ **We will assume, without loss of generality, that larger values of Y are associated with the diseased population**; If, in fact, smaller values of Y are associated with the diseased population, methods presented here could be applied by simply reversing the roles of “cases” (subjects with the disease) and “controls” (subjects without the disease).



# SENSITIVITY

- ▶ With our assumption that larger values of  $Y$  are associated with the diseased population, sensitivity  $\Pr(T=+|D=+)$ , associated with cut-point  $Y=y$  is:

$$S^+(y) = \Pr(Y > y | D=+) = \text{“true positive rate”}$$

$$= 1 - \Pr(Y \leq y | D=+) = 1 - F^+(y)$$

- ▶ where  $F^+(y) = \Pr(Y \leq y | D=+)$  is the **cumulative distribution function (cdf)** of  $Y$  for the diseased population (or population of cases).
- ▶ Sensitivity is a Survival Function

# SPECIFICITY

- ▶ With our assumption that larger values of  $Y$  are associated with the diseased population, the specificity,  $\Pr(T=-|D=-)$ , associated with cut-point  $Y=y$  is:

$$S^-(y) = \Pr(Y \leq y | D=-) = F^-(y), \text{ or}$$

$$1 - S^-(y) = 1 - F^-(y) = \text{“false positive rate”}$$

- ▶ where  $F^-(x)$  is the cumulative distribution function (cdf) of  $Y$  for the non-diseased or healthy population.
- ▶ (1-Specificity) is another Survival Function

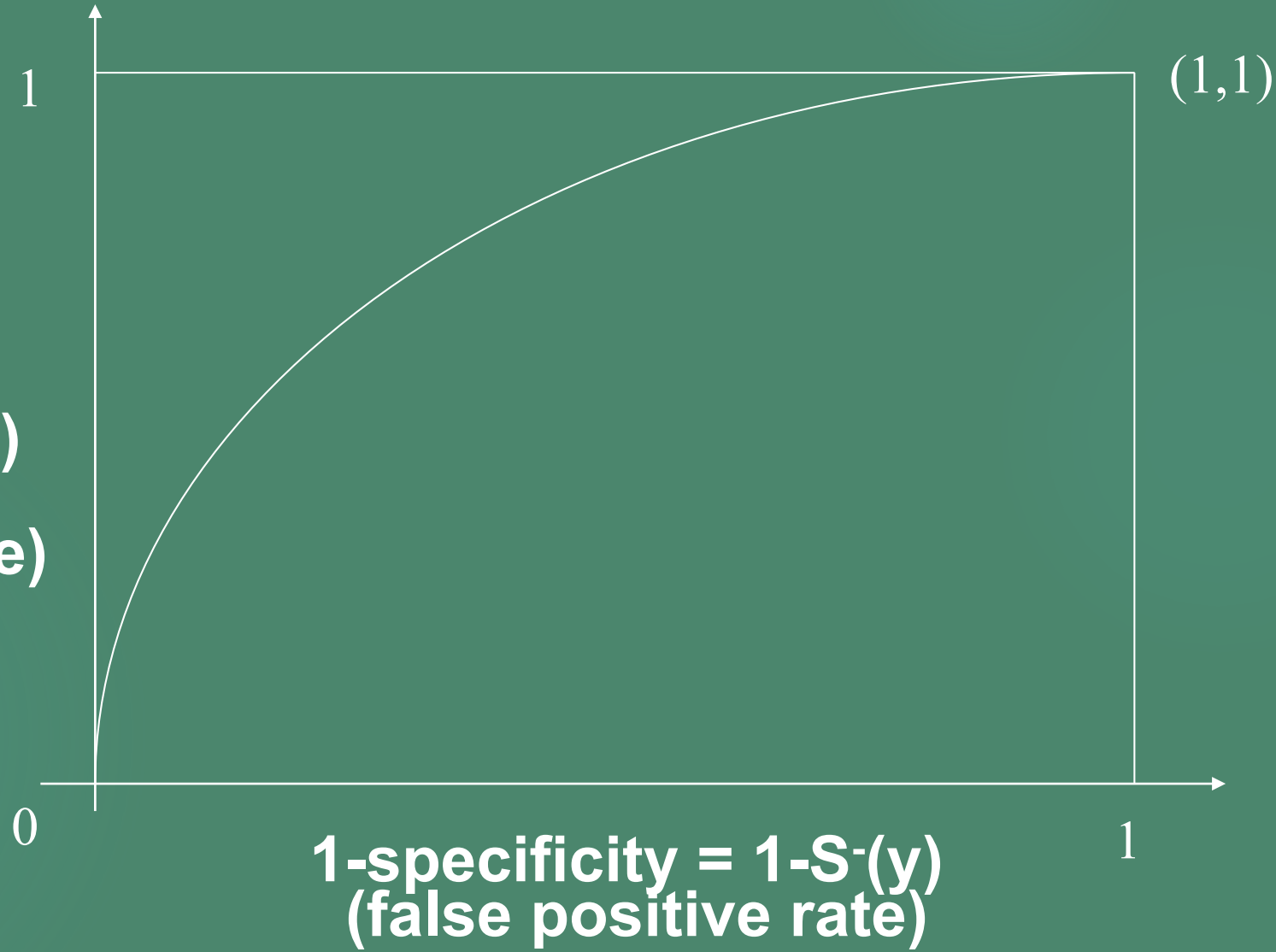
# ROC FUNCTION & ROC CURVE

- ▶ A function “R” from  $[0,1]$  to  $[0,1]$  that “maps” false positive rate (1-specificity, on horizontal axis) to true positive rate (sensitivity, on vertical axis), is called the “**ROC function**”:

$$R[1-F^-(y)] = 1-F^+(y) \text{ or } R[1-S^-(y)] = S^+(y)$$

- ▶ **The graph of  $R(\cdot)$  is called the “ROC curve”**
- ▶ The ROC curve, **the graph of sensitivity versus (1-specificity)**, is generated as the “cutpoint”  $y$  moves through its range of possible values.

**sensitivity =  $S^+(y)$**   
**(true positive rate)**



**“ROC” Curve**

The “ROC function” maps “sensitivity against (1-specificity)” or “true positive rate against false positive rate”. **It maps a survival function against another survival function.**

$$S_D(t) = 1 - F^+(t)$$

$$S_H(t) = 1 - F^-(t)$$

$$R[S_H(t)] = S_D(t)$$

$$\mathbf{R}(\mathbf{u}) = \mathbf{S}_D[\mathbf{S}_H^{-1}(\mathbf{u})]; \mathbf{u} \in [0,1]$$

# PROPORTIONAL HAZARDS MODEL

Since what we have on the axes of the ROC curve are two survival functions, one possible model is the “Proportional Hazards Model” (The only parameter is the “Hazard Ratio” or “Relative Risk”):

$$1 - F^+(y) = (1 - F^-(y))^{\theta}, \text{ or}$$

$$v = R(u) = 1 - (1 - u)^{\theta}; 0 \leq u \leq 1$$

# Index for DIAGNOSTIC ACCURACY

- ▶ ROC curve is a graphical device to show all possible combinations of sensitivity and specificity but, for simplicity, it is desirable to reduce an entire curve to a single quantitative index of diagnostic accuracy.
- ▶ Possibilities include the **difference between means of Y for the two populations divided by SD (effect size)**, those with disease and those without; and the ratio of variances. However, the most popular one has been the **area under the ROC curve**.
- ▶ The area under the curve has a powerful interpretation and it is related to other well-known statistics making it easier to learn its statistical properties.

Suppose that an observation  $y_1$  is randomly sampled from the diseased population and another random observation  $Y_0$  is independently sampled from the non-diseased population; and let  $\Pr(Y_1 > Y_0)$  denote the probability of the event that the  $Y_1$  observation is larger than the  $Y_0$  observation; we have:

$$A = \Pr(Y_1 > Y_0)$$

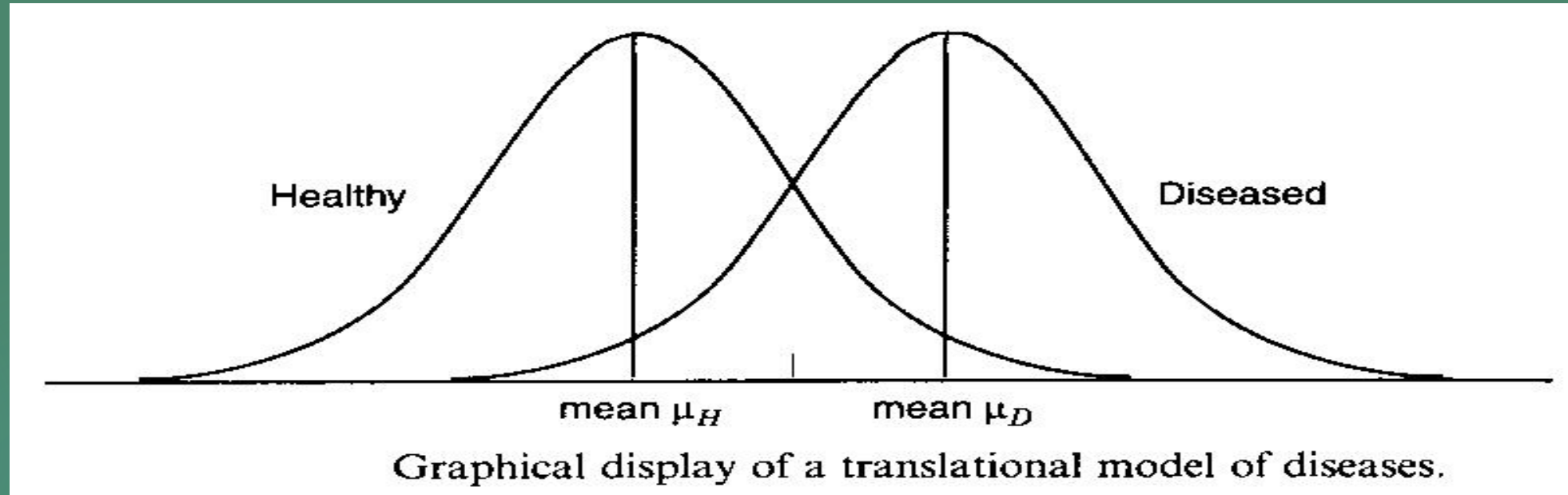
$$A = \int (1 - F^+) d(1 - F^-)$$

$$A = \int_0^1 R(u) du$$

$$A = \text{Area under ROC curve}$$



# AN ALTERNATIVE INDEX



Biomarker Y is normally distributed with the same variance, but different means; no matter where you “cut”, both errors result! The sizes of these errors depend on the “standardized distance”

$$d = (\mu_D - \mu_H) / \sigma$$

Are the two indices “A” and “d” related?

These are different numerical values but “statistically equivalent”. If we let “ $\Phi^{-1}(\cdot)$ ” denote “inverse of the standard normal cumulative distribution function”, for example  $\Phi^{-1}(.975) = 1.96$ , then Simpson and Fitter (1973) showed that:

$$d = \sqrt{2} \Phi^{-1}(A)$$

The index “d”, often called the “Effect Size”, has a very powerful interpretation in terms of disease development!

# LOGISTIC REGRESSION

The probability of disease development and the value  $Y=y$  of the biomarker  $Y$  are related by the “Logistic Regression Model”:

$$\pi_y = \Pr(\mathbf{D} = + \mid \mathbf{Y} = y) = \frac{e^{\beta_0 + \beta_1 y}}{1 + e^{\beta_0 + \beta_1 y}}, \text{ or}$$

$$\ln \frac{\pi_y}{1 - \pi_y} = \beta_0 + \beta_1 y$$

# USE OF BAYES' RULE

$$\frac{\pi_y}{1 - \pi_y} = \frac{\Pr(D = + | Y = y)}{\Pr(D = - | Y = y)}$$

$$\frac{\pi_y}{1 - \pi_y} = \frac{\Pr(Y = y | D = +) \Pr(D = +) / \Pr(Y = y)}{\Pr(Y = y | D = -) \Pr(D = -) / \Pr(Y = y)}$$

$$\frac{\pi_y}{1 - \pi_y} = \frac{\Pr(Y = y | D = +) \Pr(D = +)}{\Pr(Y = y | D = -) \Pr(D = -)}$$

$$\ln \frac{\pi_y}{1 - \pi_y} = \mathbf{Constant} + \ln \left[ \frac{\Pr(Y = y | \mathbf{D} = +)}{\Pr(Y = y | \mathbf{D} = -)} \right]$$

# RESULT

Suppose  $Y$  is normally distributed with the same variance, different means for  $\Pr(Y=y|D=+)$  and  $\Pr(Y=y|D=-)$ , we have:

$$\ln \frac{\pi_y}{1 - \pi_y} = \text{Constant} + \ln \left[ \frac{\Pr(Y = y | D = +)}{\Pr(Y = y | D = -)} \right]$$

$$\ln \frac{\pi_y}{1 - \pi_y} = \text{Constant} + \ln \left[ \frac{\exp \{ -(y - \mu_D)^2 / \sigma^2 \}}{\exp \{ -(y - \mu_H)^2 / \sigma^2 \}} \right]$$

$$\ln \frac{\pi_y}{1 - \pi_y} = \text{Constant} + \frac{(\mu_D - \mu_H)}{\sigma^2} y$$

$$\mathbf{d} = \beta_1 \sigma$$

# INTERPRETATION OF “d”

Under logistic model and Suppose Y is normally distributed with the same variance but different means for  $\Pr(Y=y|D=+)$  and  $\Pr(Y=y|D=-)$ , then:

$$d = \beta_1 \sigma$$

The value of Index “d” is equal to the log(Odds Ratio) due to a change of “one SD” in the value of the marker Y (that’s the Odds for disease development)



# **THE OPTIMIZATION PROBLEM (Dichotomization of a Biomarker)**

Diagnostic tests have been understood by patients as dichotomous outcomes but most biomarkers are on continuous scale; **PSA for prostate cancer is typical case.**

For practical application (the main objective of translational research and medicine), **the biomarker under investigation needs to be dichotomized.** After tested, the Doctor needs to tell the patient if he/she has the disease; or at least, he/she likely has the disease. One cannot call some process a “test” unless one can make a decision.



**We all know that, for example, high PSA likely indicates prostate cancer; but how high it is to classify a man as having prostate cancer? To form a diagnosis, we need to dichotomize this continuous biomarker.**

**If we set the cut-point too high, we would miss cases – that is “low sensitivity”; if we set the cut-point too low, we would have many false positives – that is “low specificity”!**


For a continuous biomarker such as “PSA”; the basic question is “How high is high?” or “How low is low?”. In practice, **cutpoints are formed arbitrarily because we fail to form and justify a criterion or criteria.**

We need an “optimal cutpoint” ; but what do we mean by “optimal”? **“Good”, but what it is good for?** May be more than one solution because there are different criteria.


**We can arrange data into a 2x2 table after a dichotomization at certain cut-point:**

	Disease	No Disease
Positive Test Result	True Positive (TP) a	False Positive (FP) b
Negative Test Result	False Negative (FN) c	True Negative (TN) d

**At this cut-point: Sensitivity  $S_+ = a/(a+c)$   
Specificity  $S_- = d/(b+d)$**



To judge the “value” of a biomarker at this cut-point, we need to **measure the strength of the relationship between Disease D and Test T**; an useful statistic must be **independent of disease prevalence** (so that it does not depend on the numbers of cases and controls in the sample which are set arbitrarily)



**There are a number of possibilities, some have been investigated elsewhere in literature; they are all expressible as functions of sensitivity and specificity. And there is a good possibility that they are all equivalent when used in the search for an optimal cut-point**

# RESPONSE DIFFERENCE (RD)

The first index is the difference (RD) between response rates from the cases and the controls – on the “additive scale”. It turns out that RD is identical to the Youden’s Index J introduced in a previous lecture.

$$\begin{aligned} \text{RD} &= \text{Pr}(T=+|D=+) - \text{Pr}(T=+|D=-) \\ &= \text{Pr}(T=+|D=+) - [1 - \text{Pr}(T=+|D=-)] \\ &= S^+ - (1 - S^-) = S^+ + S^- - 1 = J \end{aligned}$$

# RELATIVE RESPONSE (RR) & ODDS RATIO (OR)

Relative Response (RR, similar to Relative Risk) and Odds Ratio (OR) show the difference between response rates from the cases and the controls – on the multiplicative scale:

$$RR = \Pr(T=+|D=+)/\Pr(T=+|D=-)$$

$$= S+ / (1-S-)$$

$$OR = \frac{\Pr(T=+|D=+) / 1 - \Pr(T=+|D=+)}{\Pr(T=+|D=-) / 1 - \Pr(T=+|D=-)}$$

$$= \frac{(S+)(S-)}{(1-S+)(1-S-)}$$

# DICHOTOMIZATION: EMPIRICAL SOLUTION

Given 2 independent samples, sample of cases and sample of controls, each subject with a value of the biomarker

Pool the two samples and arrange in increasing order

At each midway between two data points, form a 2-by-2 table and calculate the Sensitivity  $S^+$  and Specificity  $S^-$

At each cut-point (2-by-2 table), calculate all five indices: RD, RR, OR.

Locate the cut-point corresponds to max RD, max RR, max OR – this cut-point is the optimal cut-point; as mentioned earlier, there is a good possibility that we would have the same optimal cut-point.



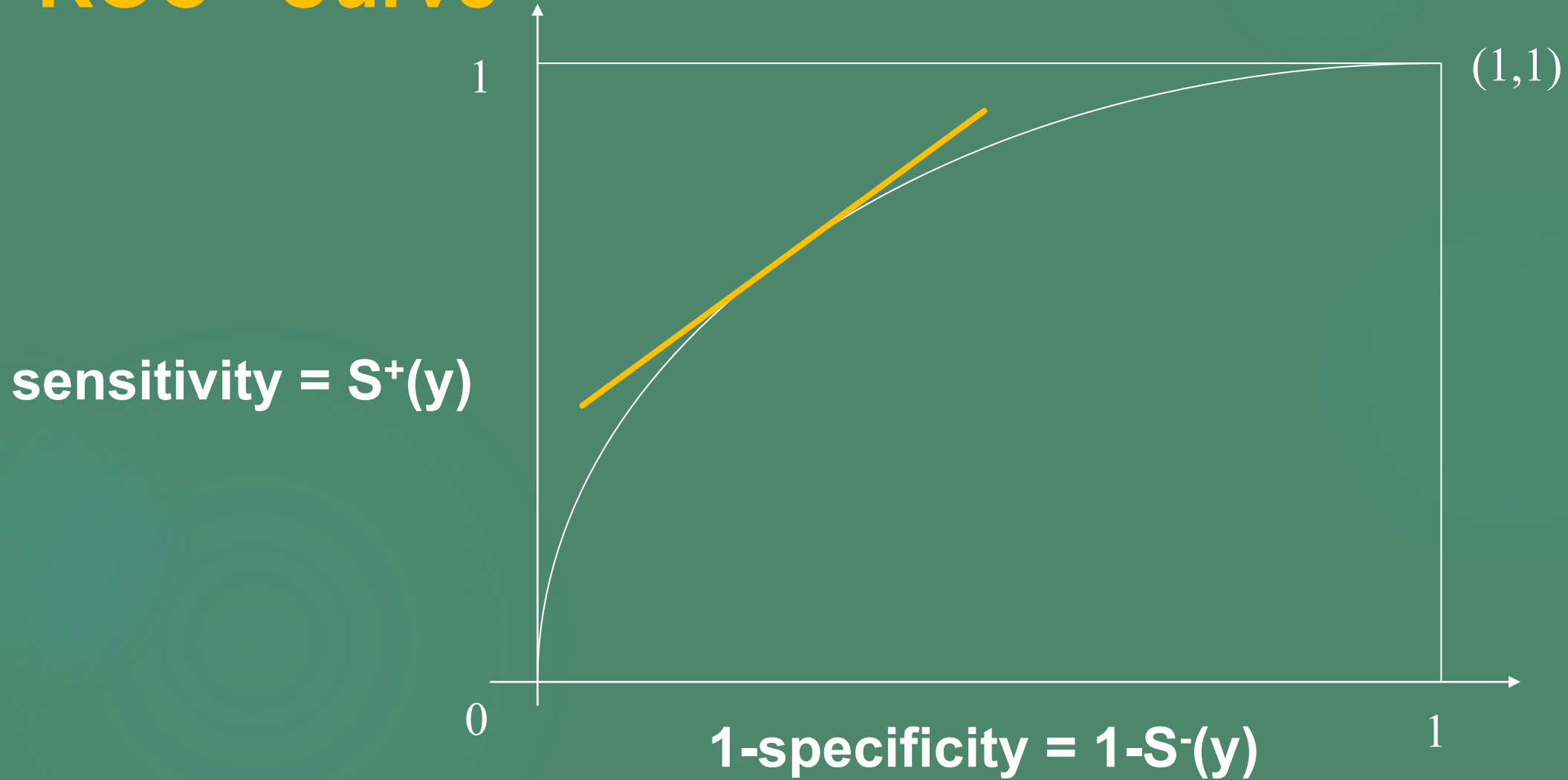
Now we have “an optimal cut-point” or maybe a few optimal cut-points if they are not identical. But can we justify or characterize the point we got; in other words what do we mean by “optimal”? **“Good”, but what it is good for?**

Let try the Response Difference,  $RD = J$ .

Since the Youden's Index ( $J = S^+ + S^- - 1 = R(U) - U$ ) is maximized when:  $0 = R'(U) - 1$ , or  $R'(U) = 1$ .

The cut-point where  $J$  is maximized has “slope = 1”

# “ROC” Curve



It is likely that, at the point where slope = 1, distance “d” to top left corner (0,1) is minimized and distance “D” to lower right corner (1,0) is maximized.

Also, recall a result from “Prevalence Survey” in which we derived **a new estimator for Disease Prevalence**

$$\pi_t = \Pr(T = +) = \Pr(T = +, D = +) + \Pr(T = +, D = -)$$

$$\pi_t = \Pr(T = + | D = +) \Pr(D = +) + \Pr(T = + | D = -) \Pr(D = -)$$

$$\pi_t = S^+ \pi + (1 - S^-)(1 - \pi)$$

$$\pi = \frac{\pi_t + S^- - 1}{J}; J = S^+ + S^- - 1, \text{ leading to}$$


$$\mathbf{p} = \frac{\mathbf{p}_t + \mathbf{S}^- - \mathbf{1}}{\mathbf{J}}$$

# STANDARD ERROR, SE(p)

$$p = \frac{p_t + S^{-1} - 1}{J}$$

$$\text{Var}(p) = \frac{\text{Var}(p_t)}{J^2}$$

$$\text{SE}(p) = \frac{1}{J} \sqrt{\frac{\mathbf{p}_t(1 - \mathbf{p}_t)}{\mathbf{n}}}$$


$$\mathbf{SE(p)} = \frac{\mathbf{1}}{\mathbf{J}} \sqrt{\frac{\mathbf{p}_t (1 - p_t)}{\mathbf{n}}}$$

Result: The “precision” of estimation of the prevalence depends only on the size of Youden’s index rather than any function of sensitivity and specificity. And this justifies the value of Youden’s index J, or Response Difference RD: **The better test is the one with larger value of the Youden’s Index or Response Difference.**



# MAXIMUM POTENTIAL OF A BIOMARKER

**“Correlation” studies relationship; it is used for Risk Determination (Risk Assessment); “Regression Analysis” is for prediction or “Diagnosis”.**

**The basic question is “How strong must the relationship be in order to have a meaningful or precise prediction?” – You cannot always say “you have lung cancer because you’re a smoker”!**

**A short answer would be “it depends on how precise you like your prediction be” or “it depends on how much error you could tolerate”**

The “ROC function” maps a survival function against another survival function, a survival function for the diseased subpopulation and a survival function for the healthy subpopulation

$$S_D(t) = 1 - F^+(t)$$

$$S_H(t) = 1 - F^-(t)$$

$$R[S_H(t)] = S_D(t)$$

$$\mathbf{R}(\mathbf{u}) = \mathbf{S}_D[\mathbf{S}_H^{-1}(\mathbf{u})]; \mathbf{u} \in [0,1]$$

To simplify the derivation, let assume a model, the same model for both subpopulations.



# LOG-LOGISTIC DISTRIBUTION

If  $\ln(X)$  is distributed as logistic,  $X$  is distributed as log-logistic; the log-logistic distribution is similar to log-normal distribution but with thicker tails – so fits better “real” non-negative measurements.

$$S(t) = \frac{1}{1 + (\rho t)^\nu};$$

$$\rho = e^{-\mu}, \text{ where } \mu \text{ is Mean}$$

$$\nu = \frac{1}{\sigma}, \text{ where } \sigma \text{ St Deviation}$$


# BOTH DISTRIBUTIONS ARE LOG-LOGISTIC

$$S(t) = \frac{1}{1 + (\rho t)^\nu}$$

$$\sigma_D = \sigma_H = \sigma$$

*Then :*

$$\begin{aligned} R(u) &= \frac{u}{u + (1 - u) \exp\left(-\frac{\mu_D - \mu_H}{\sigma}\right)} \\ &= \frac{u}{u + (1 - u)\beta} \end{aligned}$$


$$\beta = \exp\left(-\frac{\mu_D - \mu_H}{\sigma}\right)$$
$$0 < \beta < 1 \text{ for } (\mu_D > \mu_H)$$

$$R(u) = \frac{u}{u + (1-u)\beta}$$

$$R'(u) = \frac{\beta}{[u + (1-u)\beta]^2}$$

$$R'(u) = 1 \Leftrightarrow u = \frac{-\beta + \sqrt{\beta}}{1-\beta}$$

$$\text{Optimal : } S^- = 1 - u = S^+ = \frac{1 - \sqrt{\beta}}{1 - \beta}$$

$$\text{where : } \beta = \exp\left(-\frac{\mu_D - \mu_H}{\sigma}\right) = \exp(-d)$$

# MAXIMUM POTENTIAL OF A BIOMARKER


d	S <sub>-</sub> =S <sub>+</sub>
1	62%
2	73%
3	82%
4	88%

**d is called the “Effect Size”**

**Back to the case case of “Diagnosis”. We all know that, for example, high PSA likely indicates prostate cancer; but how high it is to classify a man as having prostate cancer?**

**We can see that it would take a lot to qualify as a good screening biomarker; maybe a difference of 3-4 standard deviations between cases and controls.**

# PERSONALIZED DIAGNOSIS



The approach we took only consider the relationship between the continuous biomarker and the disease status, leaving out the subjects' characteristics. **For Example, PSA is positively associated with age. Age should be incorporated to individualize the cut-point for PSA in the diagnosis of prostate cancer.** It's at an era of personalized medicine, characteristics of patients should be included to form an individualized diagnosis.





Let denote biomarker values as  $m_1, m_2, \dots, m_k$

For  $M = m_i$ ; at this cut-point, define “test”:

$T_i = 0$  (or “-”, no disease) if  $m < m_i$

$T_i = 1$  (or “+”, diseased) if  $m \geq m_i$

Let  $p = \Pr(T_i=1)$



The next step is fitting the Logistic Regression Model with 3 independent variables:  $D$ ,  $X$ , and  $D * X$  and estimate all regression coefficients:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 D * X$$

Let the estimates be  $b_0$ ,  $b_1$ ,  $b_2$ , and  $b_3$ .



Consider a specific value  $X = x$ , we have for cut-point  $m_i$ :

$$\log \frac{p}{1-p} = (b_0 + b_2x) + (b_1 + b_3x)D$$

## MODEL #1: OR-based

For cut-point  $m_i$ :

$OR_i$  = Odds Ratio relating  $T_i$  and D

$$OR_i = \exp[b_1 + b_3 x]$$

Changing cut-point from  $m_1$  to  $m_k$ , and look for the cut-point with maximum value of OR.

## MODEL #2: RD-based

For cut-point  $m_i$ , we calculate:

$$RD_i = \Pr[T_i = + | D=+] - \Pr[T_i = + | D=-]$$



Again, consider **a specific value  $X = x$** , we have at cut-point  $m_i$ :

$$\log \frac{p}{1-p} = (b_0 + b_2x) + (b_1 + b_3x)D$$

$$p = \frac{\exp[(b_0 + b_2x) + (b_1 + b_3x)D]}{1 + \exp[(b_0 + b_2x) + (b_1 + b_3x)D]}$$
$$= Pr[T = + | D]$$

For cut-point  $m_i$ :

$$RD_i = \frac{\exp[(b_0 + b_1) + (b_2 + b_3)x]}{1 + \exp[(b_0 + b_1) + (b_2 + b_3)x]} - \frac{\exp[b_0 + b_2x]}{1 + \exp[b_0 + b_2x]}$$

### Empirical Solution:

Changing cut-point from  $m_1$  to  $m_k$ , and look for the cut-point with maximum value of RD.

## Suggested Exercises:

- ▶ **#1** Refer to a dataset in file “Prostate Cancer” and suppose we focus on biomarker “Acid” in order to predict “nodal involvement”. Find a global optimal cut-point for Acid using the RD index and the optimal cut-point for subjects with positive X-ray result using the CD-based model.
- ▶ **#2** We have a data set on prostate cancer diagnosis) which includes 50 controls (subjects without prostate cancer) and 51 cases (subjects with prostate cancer); file name is “PSA-data” which was use in the last Example. Find a global optimal cut-point for PSA using the OR index and for a 65-year old subject using the OR-based regression model.