PubH 7405:

REGRESSION ANALYSIS

DISEASE DIAGNOSIS &

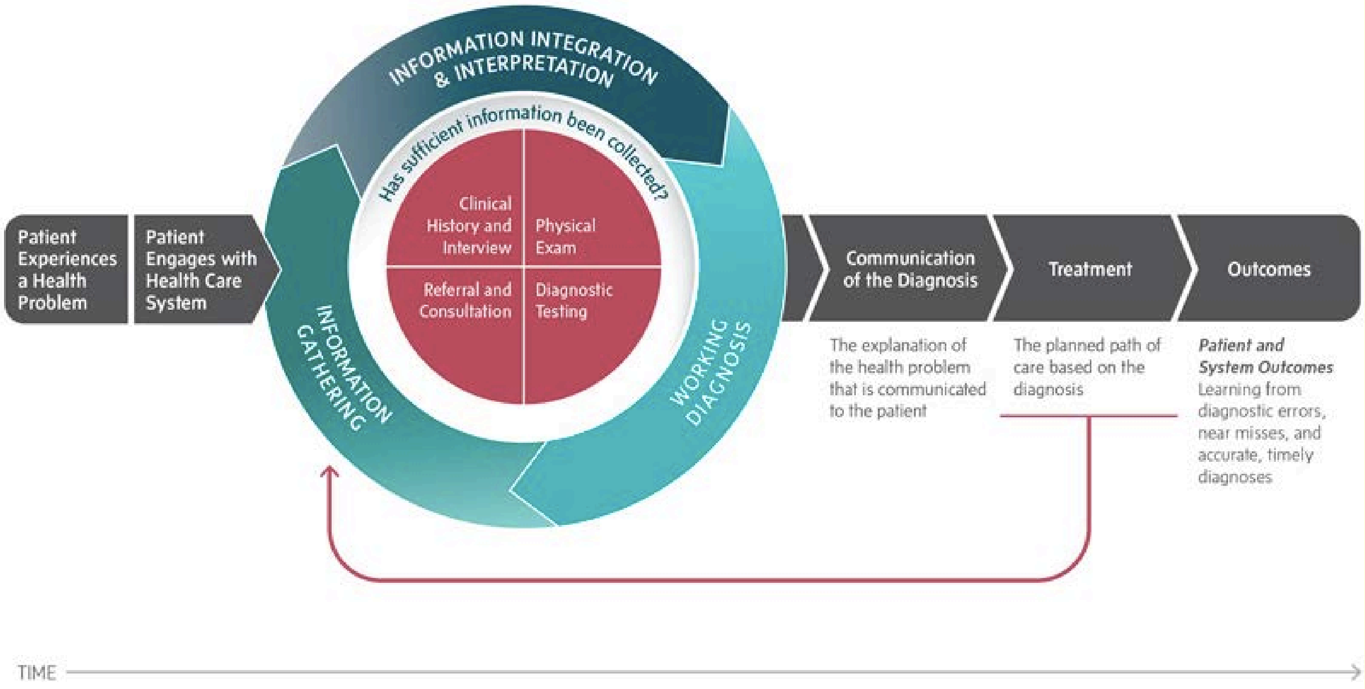
PERSONALIZED MEDICINE

- **DIAGNOSIS: The act or process of identifying or determining the nature of a disease through examination.**
- **SCREENING: The act or process of separating, or sifting out by means of an appraisal or a selection.**

DIFFERENT CONCEPTS/TERMS?

- Yes, Screening is a population-based process (Public Health) whereas Diagnosis is individually-based (Medicine).
- However, the difference is not in the make-up of the processes but in their uses.
- For the purpose of learning Biostatistics or performing data analysis, we make no strong distinction between the terms “diagnosis” and “screening”; differences, if any, are minor – two terms are often used exchangeably.

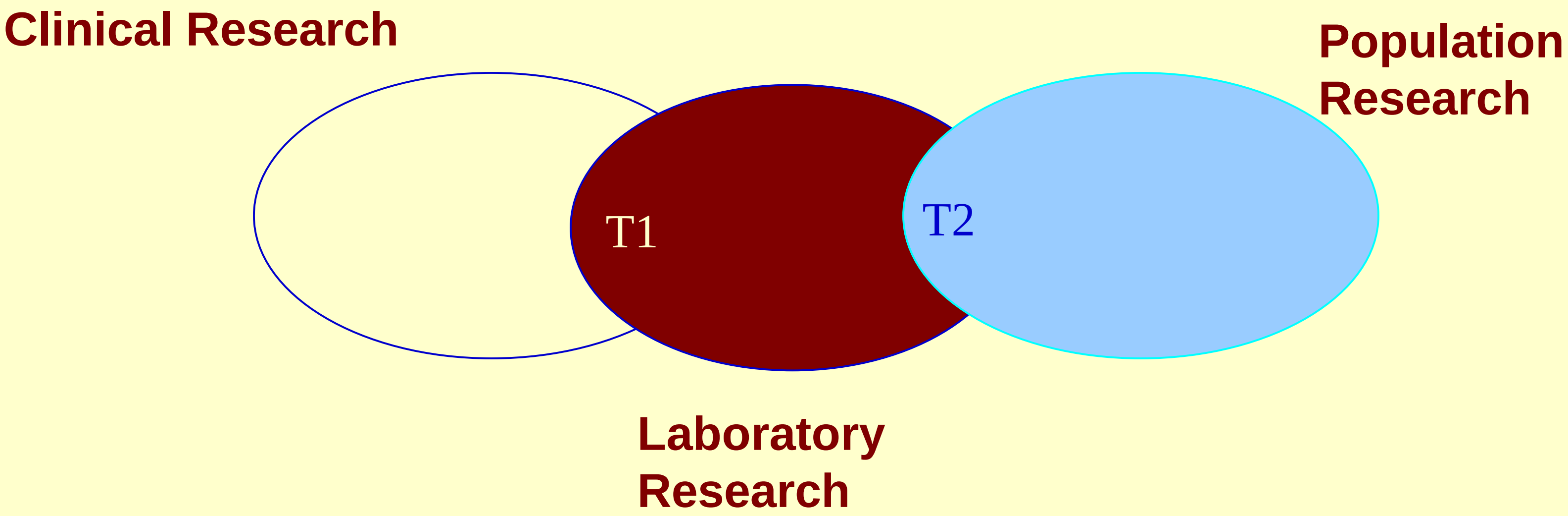
The Complete Healthcare Process


INFORMATION INTEGRATION & INTERPRETATION
Has sufficient information been collected?
Clinical History and Interview
Physical Exam
Referral and Consultation
Diagnostic Testing
INFORMATION GATHERING
WORKING DIAGNOSIS
Patient Experiences a Health Problem
Patient Engages with Health Care System
Communication of the Diagnosis
Treatment
Outcomes
The explanation of the health problem that is communicated to the patient
The planned path of care based on the diagnosis
Patient and System Outcomes
Learning from diagnostic errors, near misses, and accurate, timely diagnoses
TIME

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, a threshold, is Disease Diagnosis. It is part of Translational Research – a crossing with Basic Science. Some of us called a section of this part “Biomarker Research”.



Clinical Research

Population Research

T1

T2

Laboratory Research

Research consists of three areas: Population, Laboratory, and Clinical; Translational Research is the component of basic science that interacts with clinical science (T1) or with population science (T2). Biomarker research is in T1.

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 features: Either CONTINUOUS or ORDINAL !!!

However, for practical use/application (the main objective of translational research), 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.

A SIMPLE PLAUSIBLE MODEL

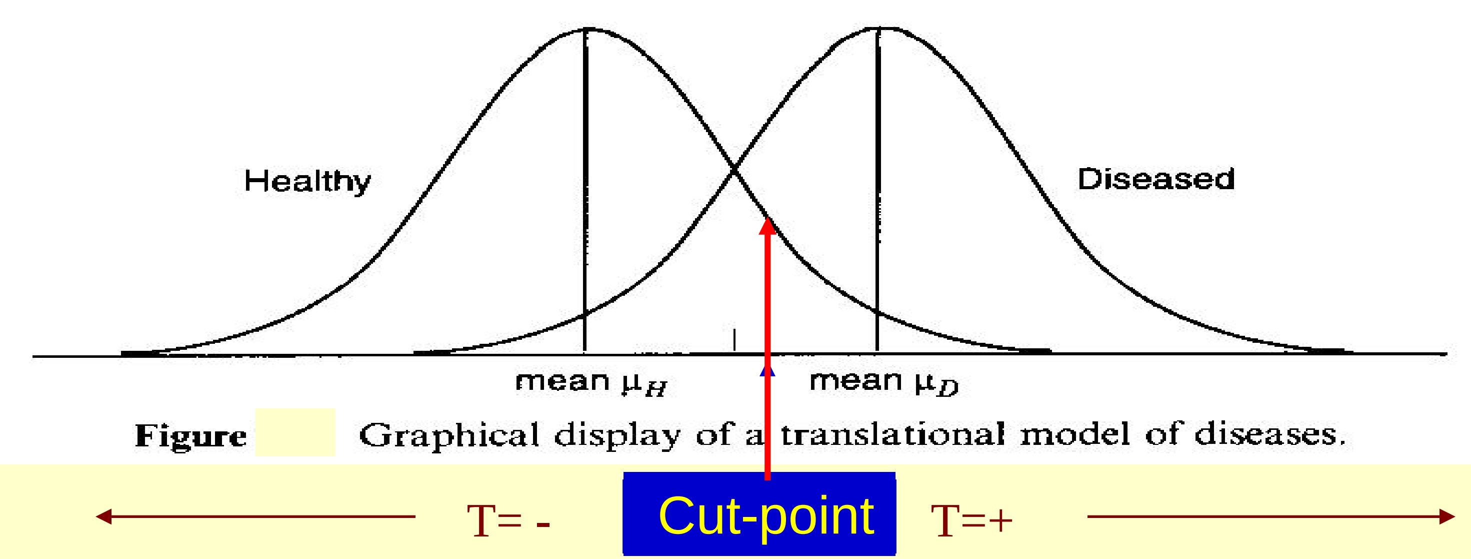
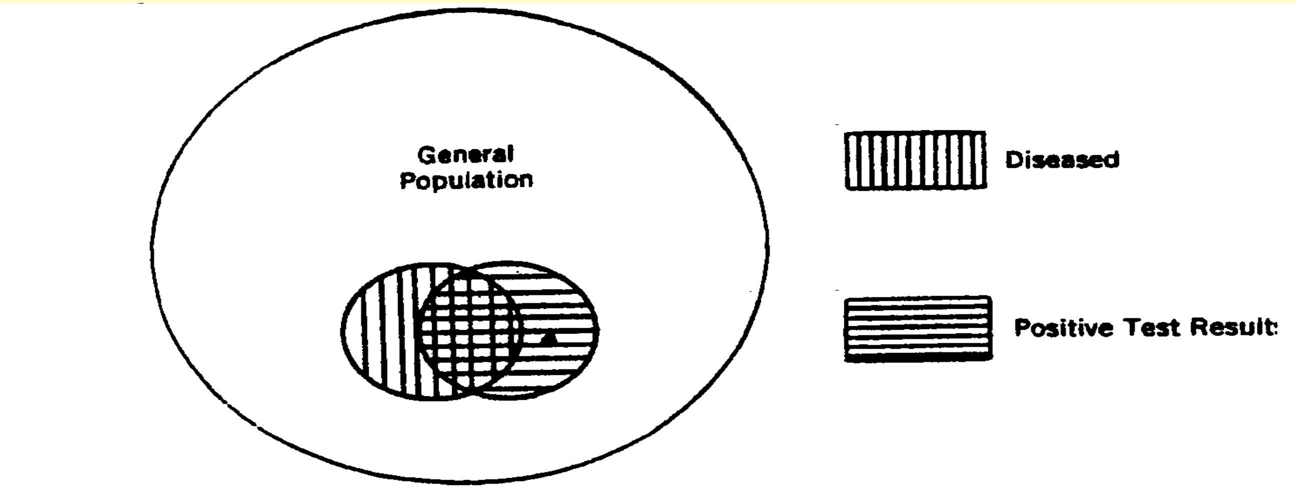


Figure Graphical display of a translational model of diseases.

Biomarker X is normally distributed with the same variance, but different means; no matter where you “cut”, both errors result!

MISCLASSIFICATION



	Test=Positive	Test=Negative
Diseased	True Positive	False Negative
Healthy	False Positive	True Negative

KEY PARAMETERS

- Let “D” and “T” denote the true diagnosis and the test result (after dichotomization), respectively
- The key parameters are two conditional probabilities):
  Sensitivity, $S^+ = \Pr(T=+|D+)$
  Specificity, $S^- = \Pr(T=-|D=-)$
- Sensitivity is the probability to correctly identify a diseased individual and Specificity the probability of correctly identify a healthy individual

PARAMETER ESTIMATION

- Sensitivity and specificity can simply be estimated as "proportions" s^+ and s^- from the two samples;
- Sensitivity is the proportion of diseased individuals detected as positive by the test; specificity is the proportion of healthy individuals detected as negative.
- Standard errors and 95% confidence intervals, for example, are calculated accordingly.

$$\text{sensitivity} = \frac{\text{number of diseased individuals who screen positive}}{\text{total number of diseased individuals}}$$

$$\text{specificity} = \frac{\text{number of healthy individuals who screen negative}}{\text{total number of healthy individuals}}$$

	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

Sensitivity $= S^{+} = P(Test = +|Disease = +) = \frac{P(Test=+,Disease=+)}{P(Disease=+)} = \frac{TP}{TP+FN} = \frac{a}{a+c}$

Specificity $= S^{-} = P(Test = -|Disease = -) = \frac{P(Test=-,Disease=-)}{P(Disease=-)} = \frac{TN}{FP+TN} = \frac{d}{b+d}$

AIDS

- Acquired Immunodeficiency syndrome (AIDS) is a severe manifestation of infection with the Human Immunodeficiency Virus (HIV, identified in 1983).
- The virus destroys the immune system leading to opportunistic infections of the lungs, brain, eyes, and other organs; Consequences include debilitating weight loss, diarrhea, and several forms of cancer.
- Currently, 40 millions living with AIDS; about 5 millions newly infected and 3 millions deaths in 2004 – most affected region is Sub-Sahara Africa. Diagnosed by blood tests.

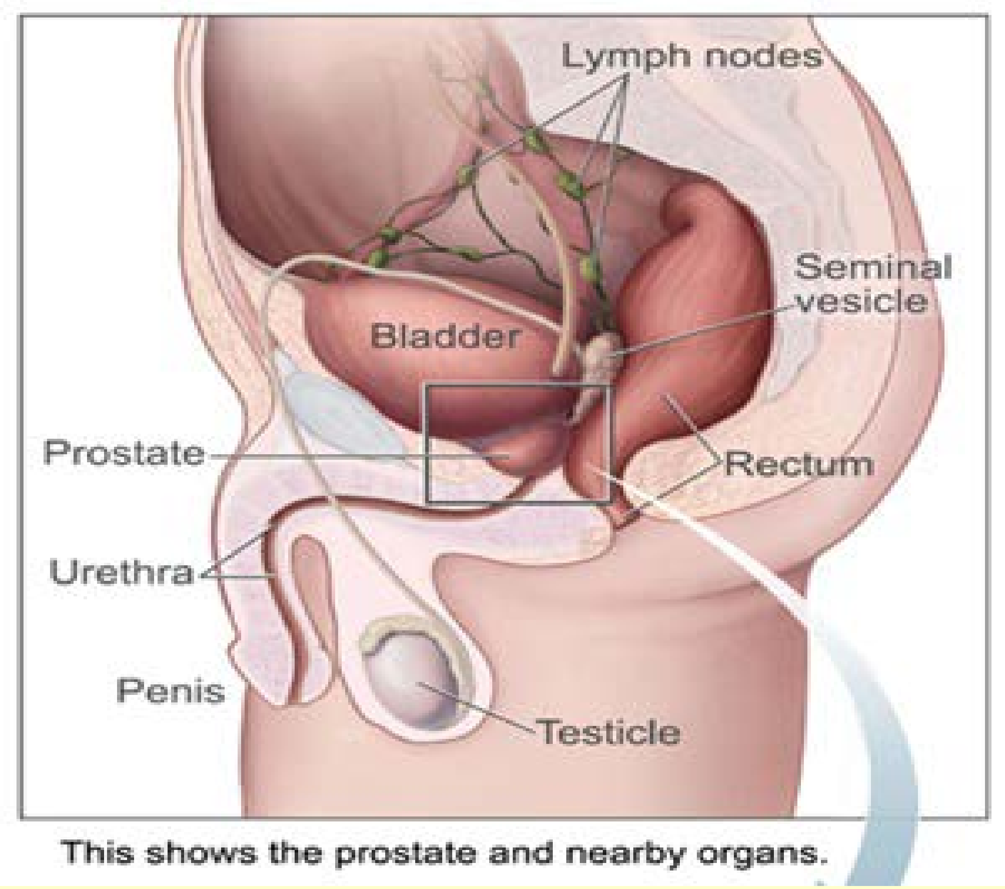
Current Estimate for USA’s AIDS prevalence is .3%

S^+=.977, S^-=.926

S^+ and S^- were determined by ELISA assay (Weiss,1985).

PROSTATE

- The prostate is part of a man’s reproductive system. It is a gland surrounding the neck of the bladder.
- A healthy prostate is about the size of a walnut and is shaped like a donut. The urethra (the tube through which urine flows) passes through the hole in the middle of that “donut”. Because of that, if the prostate grows too large, it squeezes the urethra causing a variety of urinary problems.



This shows the prostate and nearby organs.

Prostate

PROSTATE CANCER

- Cancer begins in cells, building blocks of tissues
- When normal process goes wrong, new cells form unnecessarily and old cells do not die when they should. Extra mass of cells called a tumor; and malignant tumors are cancers.
- No one knows the exact causes of prostate cancer … yet, but age is a significant factor. Most men with prostate cancer are over 65; if they live long enough a large proportion of men would eventually have prostate cancer.

PROSTATE CANCER SCREENING

- There are risk factors (age, family history) and symptoms (inability to urinate, frequent urination at night, etc…)
- Common screening is a blood test to measure prostate-specific antigen (PSA).
- However, a high level could be caused by benign prostatic hyperplasia (BPH – growth of benign cells); so the test is not very specific.

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 the absence or presence of the disease. PSA for prostate cancer is typical case.

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.

An INDEX measuring “Diagnostic Competence”

- Other things (cost, ease of application, etc…) being equal, a test with larger values of both sensitivity and specificity is obviously better.
- If not that clear cut, one has to consider the relative costs associated with 2 forms of error.
- If the 2 types of error are equally important, it may be desirable to have a single index to measure the “diagnostic competence”.

A test with larger values of both sensitivity and specificity is obviously better. But this is not always the case. 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”.

It is desirable to have one single index to measure the “diagnostic competence”. That index measures or represents the relationship between D (the disease status) and T (the test result)

	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

There are two candidates:

(1) The “Clinical Difference” (CD):

$$CD = Pr(T=+|D=+) – Pr(T=+|D=-)$$

(2) The Odds Ratio (OR)

The Clinical Difference (CD) can be expressed in a different way:

$CD = Pr(T=+|D=+) - Pr(T=+|D=-)$

$= Pr(T=+|D=+) - [1 - Pr(T=-|D=-)]$

$= Pr(T=+|D=+) + Pr(T=-|D=-) - 1$

$= S^+ + S^- - 1$ (sensitivity + specificity -1)

In Diagnostic Medicine (and Disease Screening), it is called the “Youden Index” and denoted by “J”

In an article on SMMR, Le (2006) made the case for J that, a test with maximum value of J would yield prevalence estimate with minimum standard error. Some details are as follows:

If you want to know how many percent of Minnesotans having no health insurance, you would survey n people, at random. If x of the n people in the sample have no health insurance, our estimate is x/n. This estimate, a proportion, is good – i.e. “unbiased”!

What if you want to estimate a disease prevalence? Say, what is the prevalence of HIV infection?

Or of breast cancer?

Well, you need a disease screening procedure.

But, use of a screening procedure involves errors, false positives and false negatives; so how do we estimate the disease prevalence, in light of these errors?

SETTINGS

- This is a very simple design
- We have a screening test T; its sensitivity S^+ and specificity S^- have been independently established.
- A "prevalence survey" is conducted in one target population in order to estimate the disease prevalence, $\pi = \Pr(D=+)$.
- Data: x of n subjects found "positive".

Is This a Solution?

It seems a simple solution: to estimate the disease prevalence by the frequency of positive tests: $p_t = x/n$ – ignoring its errors.

This is a good estimate but it is an estimate of $\pi_t = \Pr(T=+)$, the “response rate” whereas we want to estimate the disease prevalence, $\pi = \Pr(D=+)$. If $p_t = x/n$ is used to estimate disease prevalence, it is heavily biased upward.

A NEW POINT ESTIMATE

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

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

$$\boldsymbol{\pi}_{\mathbf{t}} = \mathbf{S}^{+}\boldsymbol{\pi} + (\mathbf{1}-\mathbf{S}^{-})(\mathbf{1}-\boldsymbol{\pi})$$

$$\boldsymbol{\pi} = \frac{\boldsymbol{\pi}_{\mathbf{t}} + \mathbf{S}^{-} - \mathbf{1}}{\mathbf{J}}; J = S^{+} + S^{-} - 1, \text{ leading to}$$

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

J is the Youden’s Index

A correction, using p instead of p_t = x/n, is a substantial improvement; in addition, if S^+ and S^- are known apriori (without errors), then p is unbiased for π.

$$\pi = \frac{\pi_t + S^- - 1}{J}; J = S^+ + S^- - 1$$

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

$$\mathbf{E(p)} = \frac{\boldsymbol{\pi}_\mathbf{t} + \mathbf{S}^- - \mathbf{1}}{\mathbf{J}}$$

$$= \boldsymbol{\pi}$$

STANDARD ERROR, SE(p)

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

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

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

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

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 is a very important result which justifies the value of Youden’s index J: The better test is the one with larger value of the Youden’s Index.**

In addition, we can easily prove that the two candidates, CD and OR, are equivalent.

	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

OR = ad/bc

	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

$$\mathbf{CD = a/(a+c) - b/(b+d)}$$

$$\mathbf{= (ad - bc)/(a+c)(b+d)}$$

OR measures the strength of the relationship on multiplicative scale, focusing on the ratio ad/bc;
CD measures the strength of the relationship on additive scale, focusing in the difference (ad-bc).

In summary, the search for an optimal cutpoint (for a continuous biomarker in order to form a disease diagnosis) would go as follows:

(1) Identifying possible cutpoints (For example, in a case-control design, these are midpoints between biomarker values);

(2) Optimal cutpoint is the cutpoint corresponding to maximum value of CD or OR.

Possible Issue

Those approaches 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.

A PSA=5 maybe too high for Tom (need a lower one to catch the disease).

Tom Age 40

Bob Age 70

A PSA = 6 maybe still low for Bob (need a higher one to recommend a biopsy).

It’s at an era of personalized medicine, characteristics of patients should be included to form an individualized diagnosis.

REGRESSION MODELS
FOR PERSONALIZED DIAGNOSIS

Data:

Case-control type consisting of Disease
(1=yes/0=no; case or control); e.g. D = Prostate Cancer
Biomarker under investigation (continuous); e.g. M = PSA
Covariates (subjects' characteristics); e.g. Age, Race, etc…, say, X_i, i=1,2,…k.
(For simplicity, consider just one covariate, X)

Aim:

Given values of covariate X, find an optimal cut-point for biomarker (to classify subjects as diseased/no-disease)

A Prototype data set (Prostate Cancer Diagnosis):
50 controls (subjects without prostate cancer)
and 51 cases (subjects with prostate cancer).
PSA values range from 0.1 to 44.6.
Data also include Age (ranging from 47 to 72);
(unfortunately, no information on Race – but
models/methods could handle more factors)

Let denote biomarker values as m_1, m_2, …, m_k

For M = mi; at this cut-point, define “test”:
T_i = 0 (or “-”, no disease) if m<mi
T_i = 1 (or “+”, diseased) if m>=mi

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:

$$log\frac{p}{1-p} = (\mathbf{b_0} + \mathbf{b2x}) + (\mathbf{b_1} + \mathbf{b3x})\mathbf{D}$$

MODEL #1: OR-based

For cut-point m_i:

ORi = Odds Ratio relating T_i and D

OR_i = exp[b1+b3x]

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

MODEL #2: CD-based

For cut-point m_i, we calculate:

CD_i = Pr[T_i = +| D=+] – Pr[T_i =+ | D=-]

Again, consider a specific value X = x, we have for cut-point $\mathrm{m_i}$:

$$\boldsymbol{log\frac{p}{1-p}} = (\mathbf{b_0} + \mathbf{b2x}) + (\mathbf{b_1} + \mathbf{b3x})\mathbf{D}$$

$$\mathbf{p} = \frac{\boldsymbol{exp[(b_0+b_2x)+(b_1+b_3x)D}}{\boldsymbol{1+exp[(b_o+b_2x)+(b_1+b_3x)D]}}$$

$$= \boldsymbol{Pr[T = +|D]}$$

For cut-point m_i:

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

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

Due As Homework

- #24.1 Refer back to the dataset “Prostate Cancer” (used the lecture on Logistic Regression”.; and suppose we focus on biomarker “Acid” in order to predict “nodal involvement”. Find the global optimal cut-point for Acid (no other information used” and the optimal cut-point for subjects with positive X-ray result using the CD-based model.
- #24.2 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 optimal cut-point for PSA for a 65-year old subject using the OR-based model.