Data Management and Analysis

Prof. Dr. San San Htay
Prof. and Head
Department of Preventive and Social Medicine
University of Medicine 1, Yangon

R question

- Research question
- Research hypothesis

Objective

- General objective
- Specific objectives
- Statistical hypothesis

Research methodology

- Study design,
- area, period
- Study population
- Sample size and sampling
- **Data collection**
- **Data management, and analysis**

Data management process

- Data collection
- Data cleaning and editing by manual
- Data compilation in data sheet
- Data entry by computer and statistics software
- Data editing (checking, correcting and recoding)
- **Data summarization**
- **Data analysis**
- **Data interpretation**

Epidemiological Study/ Designs/ Methods / Strategies

A. Observational / non - intervention / non-experimental studies

1. Descriptive study

- Case-study
- Case series
- Cross sectional (descriptive, prevalence study,)

2. Analytic study

- Ecological or correlation studies
- Cross sectional (analytic/comparative)
- Case-control study
- Cohort study (prospective)

B. Experimental / intervention / non-observational Studies

- RCT (Randomized Controlled Trial)
  - Randomized controlled clinical trial
  - Field trial,
  - Community trial
- Quasi-experimental
  - before and after intervention,
  - non randomized control trial

Statistics deals with the techniques for data

- Collection
- Summarising
- Analysis
- Interpretation

Scales of measurements

- Nominal: Distinguished / mutually exclusive
- Ordinal : Distinguished and ranked
- Interval: Distinguished, ranked & equal distant
- Ratio: Distinguished, ranked, constant units & true zero

Type of Data/ Variables

Categorical (Qualitative)

names or labels

NOMINAL

-Categories
-mutually exclusive,
-un-order

ORDINAL

-Categories are order

Numerical (Quantitative)

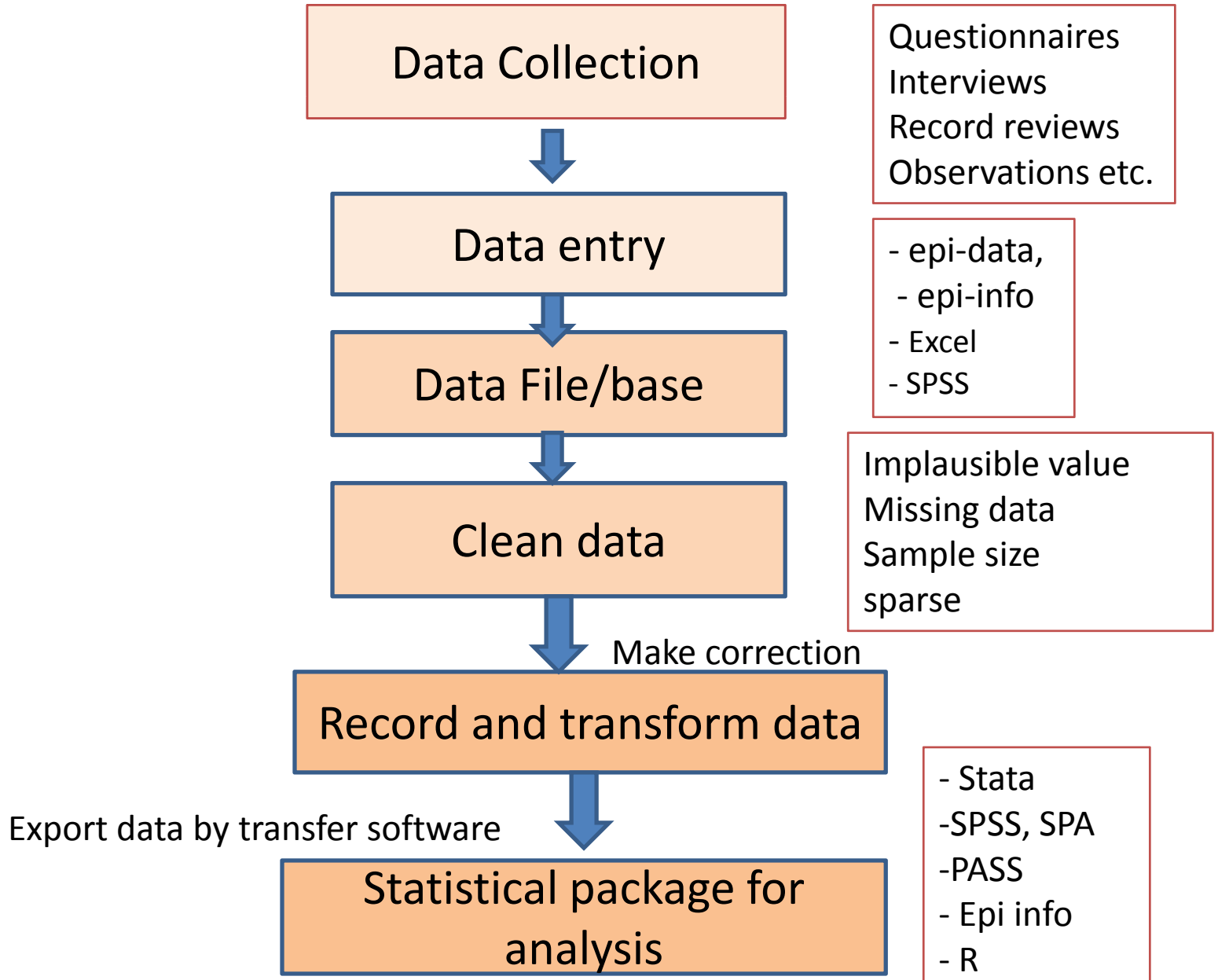
measurable quantity

DISCRETE

-Integer values,
-typically counts
-Gap, Interruption

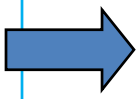
CONTINUOUS

-any value in a range of value
- measure
-no gap

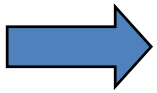

Data Collection
Questionnaires
Interviews
Record reviews
Observations etc.
Data entry
- epi-data,
- epi-info
- Excel
- SPSS
Data File/base
Clean data
Implausible value
Missing data
Sample size
sparse
Make correction
Record and transform data
Export data by transfer software
Statistical package for analysis
- Stata
-SPSS, SPA
-PASS
- Epi info
- R

Data to Information

- Accuracy
- Timeliness
- Completeness
- Consistency
- Conciseness
- Relevancy



Quality data



Quality research

Validation of data

What are to be checked?

- Range
- Missing values
- Error
- Incompleteness
- Misclassification
- Inconsistencies
- Outliers
- Illogical entries

Reducing errors in data entry

- Manuals
- Training, certification
- Double entry
- Design a standard procedure for checking
- Check file

Reliability

- **Internal consistency**
  - correlations amongst multiple items in a factor
  - e.g., Cronbach's Alpha (α)
- **Test-retest reliability**
  - correlation between time 1 & time 2
  - e.g., Product-moment correlation (*r*)

Different types of Bias(error)

- Selection bias
  - Error due to systematic differences in characteristics between those who are selected for study and those who are not.
- Information (misclassification) bias
  - A flaw in measuring exposure or outcome data that results in different quality (accuracy) of information between comparison groups.
- Confounding
  - The distortion of the association between an exposure and disease outcome by an extraneous, third variable called a confounder.

Correction of confounding

Study design

- selection
- matching
- randomization

Analysis

- stratification (classification)
- standardization (adjustment)
- advanced statistical methods (multivariate analysis, logistic regression, etc.)

Data summarization
(for Continuous data)

- **Central tendency**: Mean, Median, Mode

- **Dispersion**: Standard deviation, Variance, Range, Standard error
  Inter Quartile Range (IQR)
  Coefficient of Variation (CV)

Data summarization
(for discrete data)

Basic tool

- Proportion, probability, percentage
- Ratio
- Rate

Morbidity: Prevalence and Incidence

Mortality: Mortality rate

Others: success rate, recovery, survival rate, failure , remission rate

Frequency (Measurement of disease occurrence
(Prevalence and Incidence)

Prevalence

- Definition: Prevalence is the proportion of the population that has the disease at a certain point in time or over a period of time.
- It refers to all current cases (old & new) existing at a given point in time (point prevalence) or over a period of time (period prevalence) in a given population.

$$P = \frac{\text{all current cases (old \& new)}}{\text{Study population}} * 1000$$

Incidence

- Definition: Incidence is the number of new cases occurring in a defined population during a specified period of time.
- *A. Cumulative incidence: it is the proportion of individual in disease free state at the* beginning of the period that move to the disease state during the period

CI = new cases/ population at risk

- *B. Incidence rate (Incidence density)*

I R = new cases/ sum of person time

Statistical tests for analysis

- **Z** test (for large sample: population)
- **t** tests (Student t test and Paired t test)
- ANOVA, MANOVA
- Correlation and Regression
- Chi Square test,
- Non Parametric Tests

Tests of statistical significance

‘t’ test (Student’ t)

• To compares **two** independent samples drawn from the same population

- Any difference between **two** population **means**?
- Eg. Blood glucose level between male and female

	BMI (mean, SD)	P value (t test)
Male	24±4.2	0.01
Female	26±5.2	

‘t’ test (Paired t)

- To compares two sets of observations on a **single** sample
- **Before and after clinical trial (therapeutic or preventive)**
- Difference between before mean and after mean
- Eg. Blood glucose level before and after exercise

	A1c level (mean± SD)	P value (Pair t test)
Before exercise	7.3 ±2.3	0.02
After exercise	6.4±2.1	

Chi-square test

- To analyze the **association of categorical** variables
- To check **statistically significant association** between **two** variables
- Eg. Smoking and Ca lung
- Time series trial

	HIV+	HIV-	p value (Chi sq)
TB+	60	30	0.005
TB-	40	70	

ANOVA: analysis of variance(F test)

- To compare **more than two** means
- more appropriate than multiple ‘t’ tests
- to test the null hypothesis that three or more treatments are equally effective.

	Group1 normal	Group2 Overweight	Group3 obese	F test (p value)
TC(M±SD)	180±10	240±15	280±20	.003

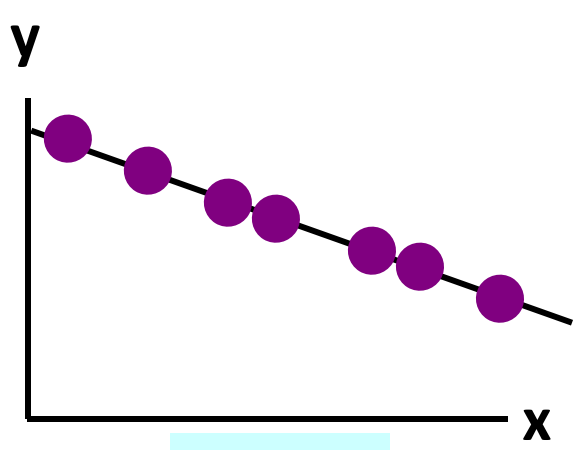
Correlation analysis

- Analyzing association/correlation of **continuous** data
- measure the **strength** of the association between two variables;
- Pearson's product moment correlation coefficient ***'r'***

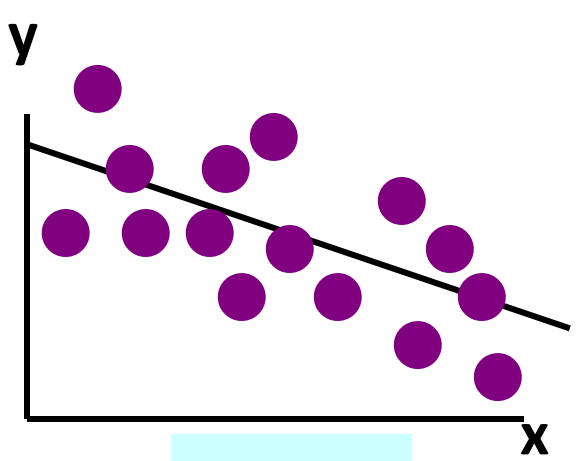
Regression analysis

- derive a **prediction equation** for estimating the value of one variable given the value of the other

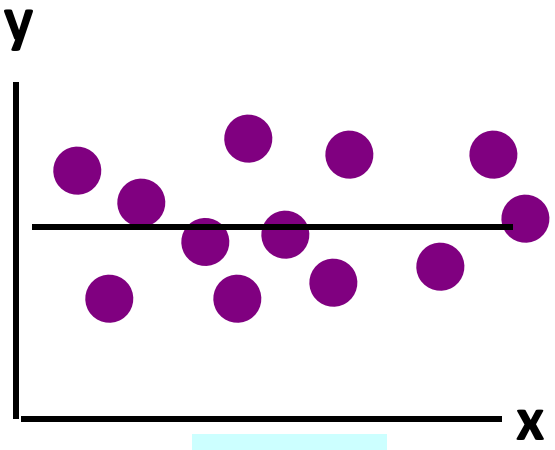
Examples of Approximate r Values



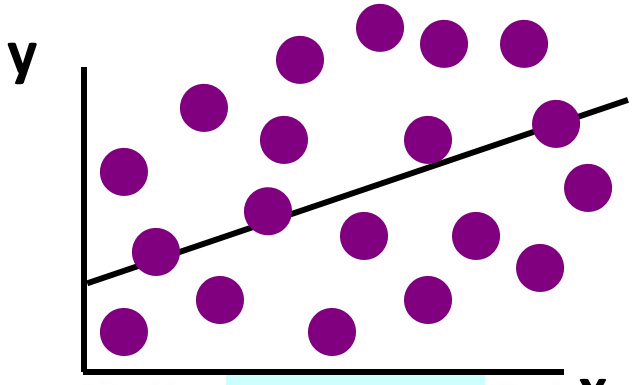
r = -1



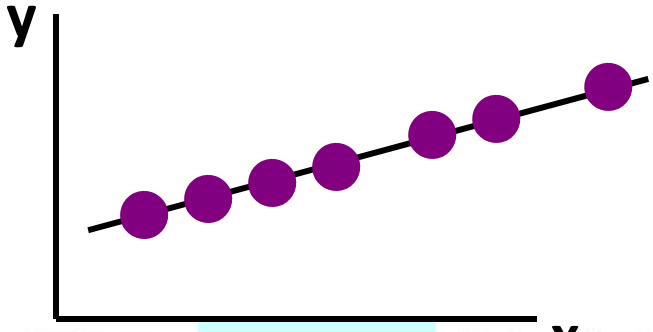
r = -.6



r = 0



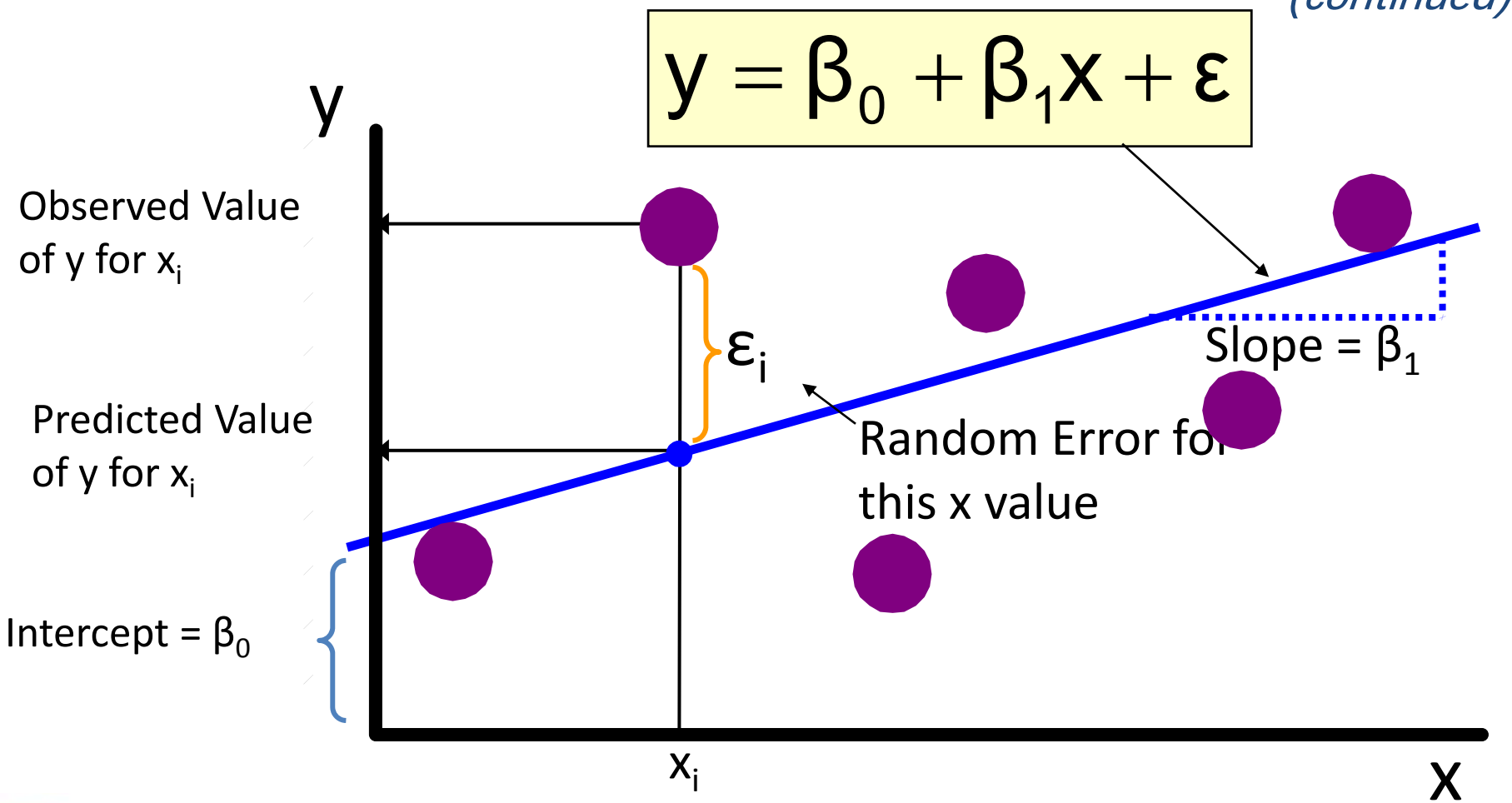
r = +.3



r = +1

Population Linear Regression

(continued)



Non Parametric Test (commonly used)

Purpose of test	Parametric test	NPT	Examples
Compares two independent samples drawn from the same population	Two sample (unpaired) *t* test	Mann-Whitney U test	To compare girls' heights with boys' heights
Compares two sets of observations on a single sample	One sample (paired) *t* test	Wilcoxon matched pairs test	To compare weight of infants before and after a feed
Assesses the strength of the straight line association between two continuous variables.	Product moment correlation coefficient (Pearson's *r*)	Spearman's rank correlation coefficient	To assess whether and to what extent plasma HbA1 concentration is related to plasma triglyceride concentration in diabetic patients
more than three groups	**One way ANOVA**	Kruskal-waillis	To determine whether plasma glucose level is higher one hour, two hours, or three hours after a meal
Association between two independent variables	Chi^2 test	Fisher's exact test	Association between smoking and lung cancer

Analytical strategies according to epidemiological study design

Cross Sectional Study

- Frequency distribution table (eg. predictor, severity, outcome etc.)

Variable	Freq,	%
Normal	50	50 %
High BP	30	30 %
Hypertension	20	20%
Total	100	100 %

Proportion of hypertension among DM pts = 20%

Cross Sectional Analytic (Comparative) Study

- Analysis for association (predator/risk and disease)

	DM+	DM-	
Obesity	a	b	a + b
Non obese	c	d	c + d
	a+c	b+d	a +
			b+c+d

Odds Ratio =ad/bc

Case control study

Analysis for association between disease and exposure

	D/s present	D/s absent
Exposed	**a**	**b**
Non- exposed	**c**	**d**
Total	**a+c**	**b+d**

To estimate the related risk - OR

OR = ad / bc

OR=2

Interpretation: Odds of developing ca lung among smokers compared to non smoker is 2.

Conclusion: Smoking is positively related with developing of ca lung.

Cohort study

Exposure Effect analysis

	D/s present	D/s absent	Total	incidence
exposed	**a**	**b**	**a + b**	**a/a+b**
non-exposed	**c**	**d**	**c + d**	**c/c+d**

Relative Risk = Incidence among exposed

(Risk Ratio) Incidence among non-exposed

$$RR = (I_E / I_{NE}) = \frac{a / a+b}{c / c+d}$$

$$AR = I_E - I_{NE} \qquad AR\% = I_E - I_{NE} / I_E * 100$$

Experimental study
(Clinical trial)

Outcome

- measurements include both improvement (the desired effect) and any side effects that may appear.

Clinical outcomes

Efficacy/effectiveness

- Outcome: cure rate /improved
- survival rate(disease free SR, 5 years SR)
- Remission rate
- Recurrent rate
- Analgesia requirement
- Hospital stay/ convalescent days
- Operation time
-

Safety

-Complication

-mortality/Death rate

-Blood transfusion

-Conversion to surgery/ re-intervention

Efficiency: cost, benefit

Efficacy& Effectiveness

- **Efficacy**, or how well a treatment works under

"ideal" conditions, may be differentiated from

- **Effectiveness**, or how well a treatment works in "real-life" situations.

Although randomized trials most often evaluate efficacy of a treatment, the two terms (efficacious and effective) are often used interchangeably.

- **Efficiency**
- cost benefit ratio
- a cheaper and better way

Randomized clinical Trial
(Analysis for effectiveness/efficacy of **Treatment**)

	Improved/ cure	Not improved	Total	Incidence
Treatment group	**a**	**b**	**a + b**	a/a+b
Placebo group	**c**	**d**	**c + d**	c/c+d

Cure rate (incidence) among treatment= I_t = a/a+b

Cure rate (incidence) among placebo= I_p = c/c+d,

AR= (I_t - I_p)

AR % = (I_t - I_p) /I_t *100

Randomized clinical Trial

(**Preventive trial**)

Assessment of efficacy of Vaccine

	Disease occurs	No disease	
Placebo	**a**	**b**	**a + b**
Treatment/Vaccine	**c**	**d**	**c + d**

Risk(disease development) among placebo= I_p = a/a+b
Risk(disease development) among vaccine = I_v = c/c+d
Risk Reduction = (I_p - I_v)
Relative Risk Reduction (%) = (I_p - Iv) /I_p *100

Reduction of Maternal-Infant Transmission of HIV with Zidovudine Treatment

	Infant HIV infected		
	YES	No	Total
Zidovudine	13	167	180
Placebo	40	143	183
Total	53	310	363

CI Ratio = (13/180)/(40/183)= 7.2%/21%= 0.33

Prevented Fraction =1 - 0.33 = 0.67= 67%

CI Diff. = (13/180) - (40/183)= 7.2%-21%=-14.1

Reduction of Maternal-Infant Transmission of HIV with Zidovudine Treatment

CI Ratio = 0.33

Interpretation: Infants whose mothers took zidovuine had one third the risk of becoming HIV infected than did the infants whose mothers took placebo.

Prevented Fraction(PF) = 67%

Interpretation: There is a 67% reduction in risk of HIV infection to infants in zidovuine group

CI Difference = -14.1%

Interpretation: 14 HIV infections among every 100 infants whose mothers took placebo would have been prevented if these mothers took zidovuine .

Ways Of Expressing The Results Of Randomized Trials

- The results of randomized trials can be expressed in a number of ways.

The risks of death or of developing a disease or complication in each group can be calculated, and **the *reduction in risk (efficacy)*** *can* then be calculated.

Efficacy of an agent being tested, such as a vaccine, can be expressed in terms of the rates of developing disease in the vaccine and placebo groups:

$$\text{Efficacy} = \frac{\left(\begin{array}{c}\text{Rate in those who}\\ \text{received the placebo}\end{array}\right) - \left(\begin{array}{c}\text{Rate in those who}\\ \text{received the vaccine}\end{array}\right)}{\text{Rate in those who received the placebo}}$$

This formula tells us the extent of the reduction in disease by use of the vaccine. Risks are often calculated per *person-years of observation (CI).*

Randomized clinical Trial (Treatment)

Assessment of bad outcomes

	Event /disease	No event/No disease	
Placebo	**a**	**b**	**a + b**
Treatment	**c**	**d**	**c + d**

Risk (bad outcomes) among placebo= I_p = a/a+b

Risk (bad outcomes) among placebo = I_t = c/c+d

Risk Reduction = (I_p - I_t)

RRR % = (I_p - I_t) /I_p *100

Number needed to treat (NNT)
Number needed to harm (NNH)

Number needed to treat (NNT)

- Number needed to be treated to prevent one more event

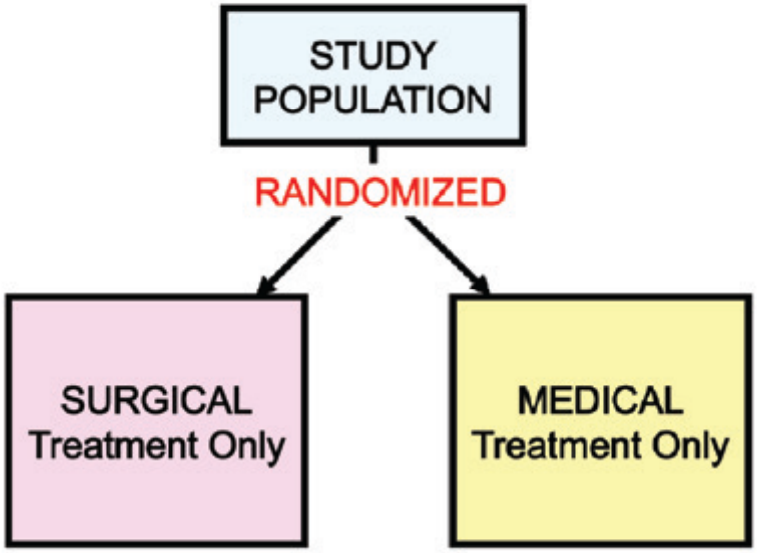
$$NNT = 1/ (Risk_{control} - Risk_{treatment})$$
$$= 1/ARR$$

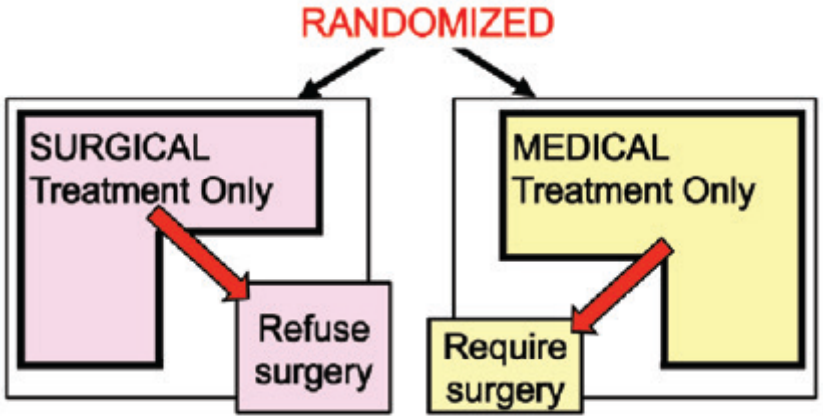
Number needed to harm (NNH)

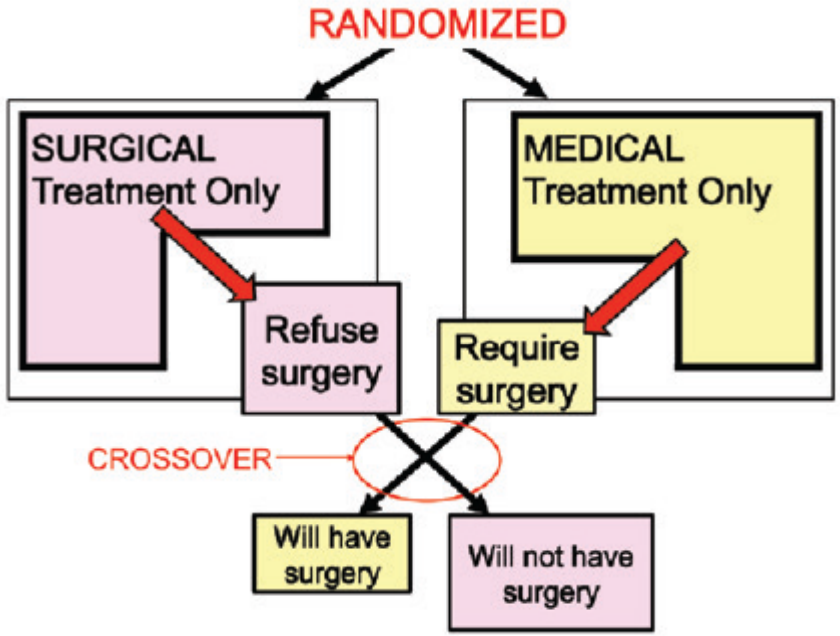
- Number needed to be treated to harm one more of them

$$NNH = 1/ (R_{treatment} - R_{control})$$

Unplanned crossover







Unplanned crossover

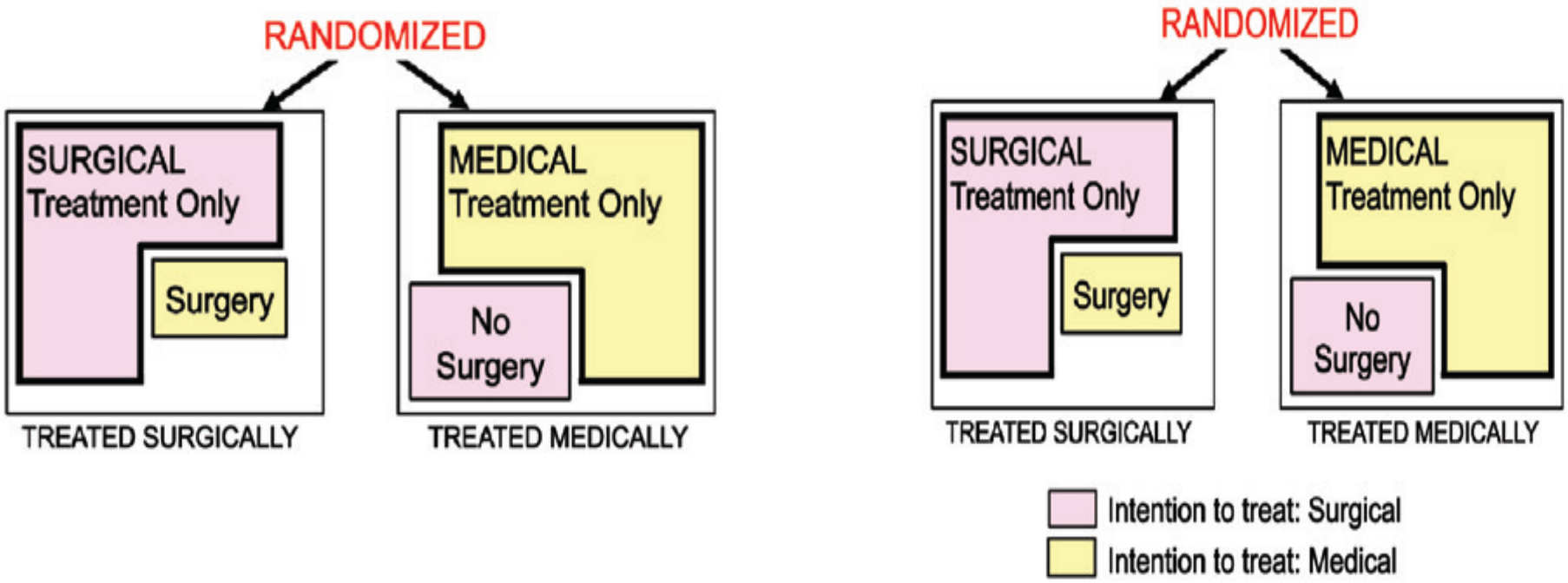


Fig. 10.6 (A–E) Unplanned crossover in a study of cardiac bypass surgery and the use of intention to treat analysis. (A) Original study design. (B–D) Unplanned crossovers. (E) Use of intention to treat analysis.

- ***Intention to treat analysis:*** If we analyze according to the original assignment

- **Per protocol analysis: that** include only participants who adhered to the protocol
- ***as treated analysis****:* if we analyze according to the treatment that the patients actually receive

Diagnostic Study

Validity (Accuracy)

-Ability of test to distinguish between who has a disease and who does not. (Sensitivity & Specificity)

Reliability (Precision, Reproducibility, Repeatability,)

- Ability of a measurement to give the same result or a very similar result with repeated measurement

2 by 2 table comparing test results and the true disease status

		True disease status	
		Diseased	**Not diseased**
Test Result	**Positive**	**True Positive (a)**	*False Positive (b)*
	Negative	*False Negative (c)*	**True Negative (d)**

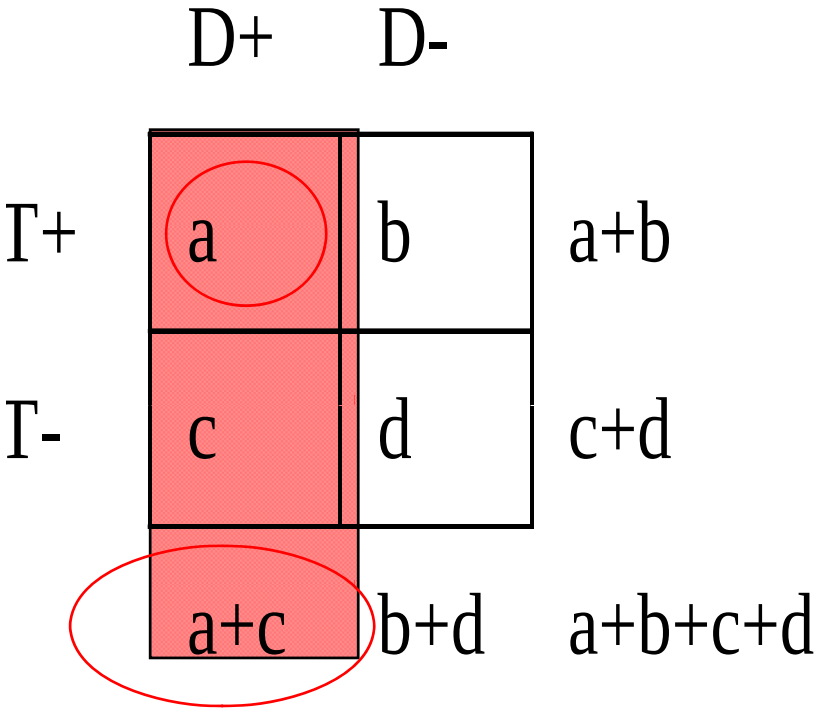
Sensitivity & Specificity

Sensitivity: ability of a test to detect a disease when it is present [a / a+c]

False negative error rate [c/a+c] = 1- sen

Specificity: ability of a test to indicate non-disease when it is absent [d/b+d]

False positive error rate [b/b+d] = 1-sp



Predictive Values

If a patient's result is positive, what is the probability that he or she has the disease being tested?

- **Positive predictive value (PPV)**: proportion of subjects who had positive test results had the disease

  [a / (a+b)]

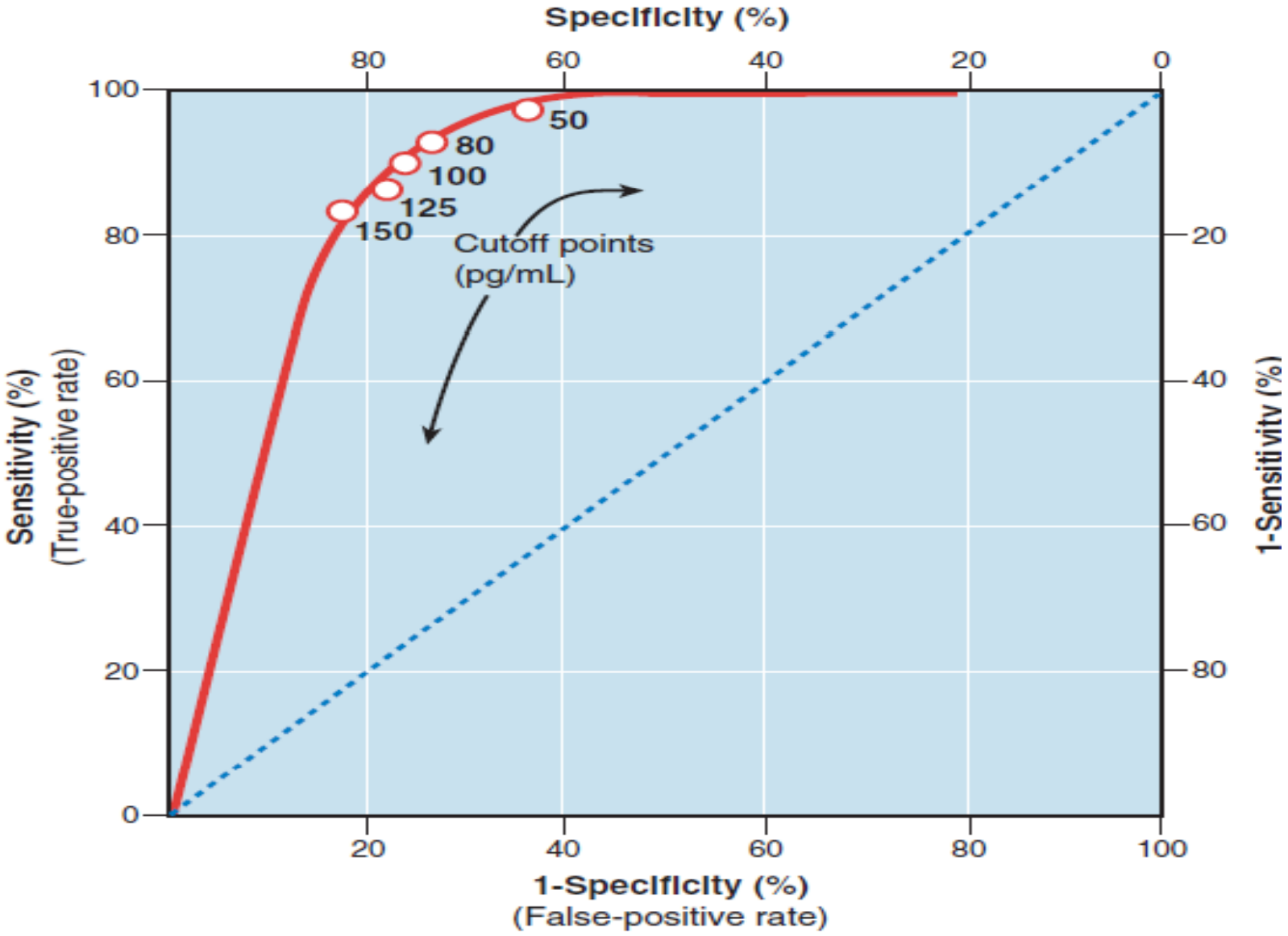
- **Negative predictive value (NPV):** proportion of subjects who had negative test results were free of the disease

  [d/(c+d)]

	D+	D-	
T+	a	b	a+b
T-	c	d	c+d
	a+c	b+d	a+b+c+d

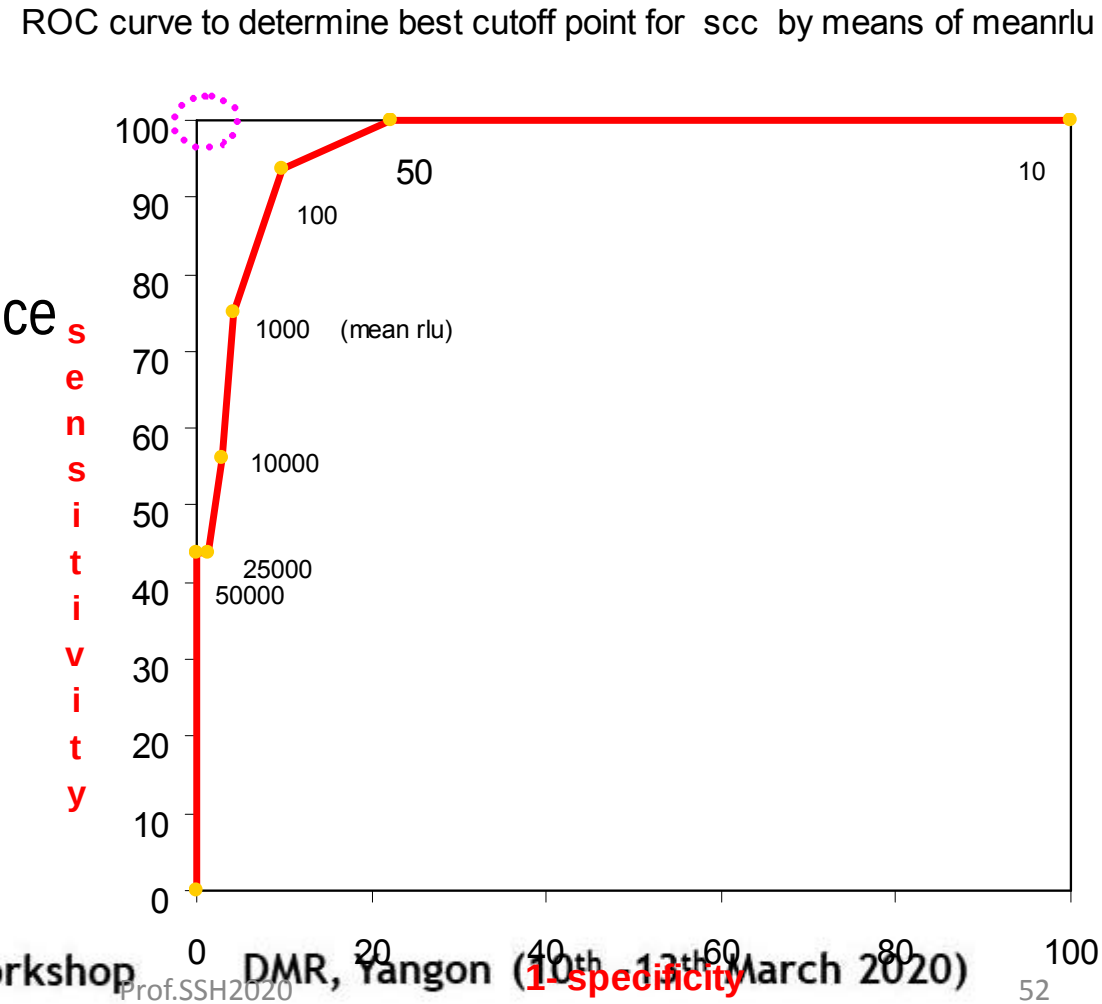
Trade-Off between Sensitivity and Specificity When Using BNP Levels to Diagnose Congestive Heart Failure

BNP Level (ph/mL)	Sensitivity (%)	Specificity (%)
50	97	62
80	93	74
100	90	76
125	87	79
150	85	83


Specificity (%)
80
60
40
20
0
100
80
60
40
20
0
50
80
100
125
150
Cutoff points
(pg/mL)
Sensitivity (%)
(True-positive rate)
1-Sensitivity (%)
20
40
60
80
100
1-Specificity (%)
(False-positive rate)

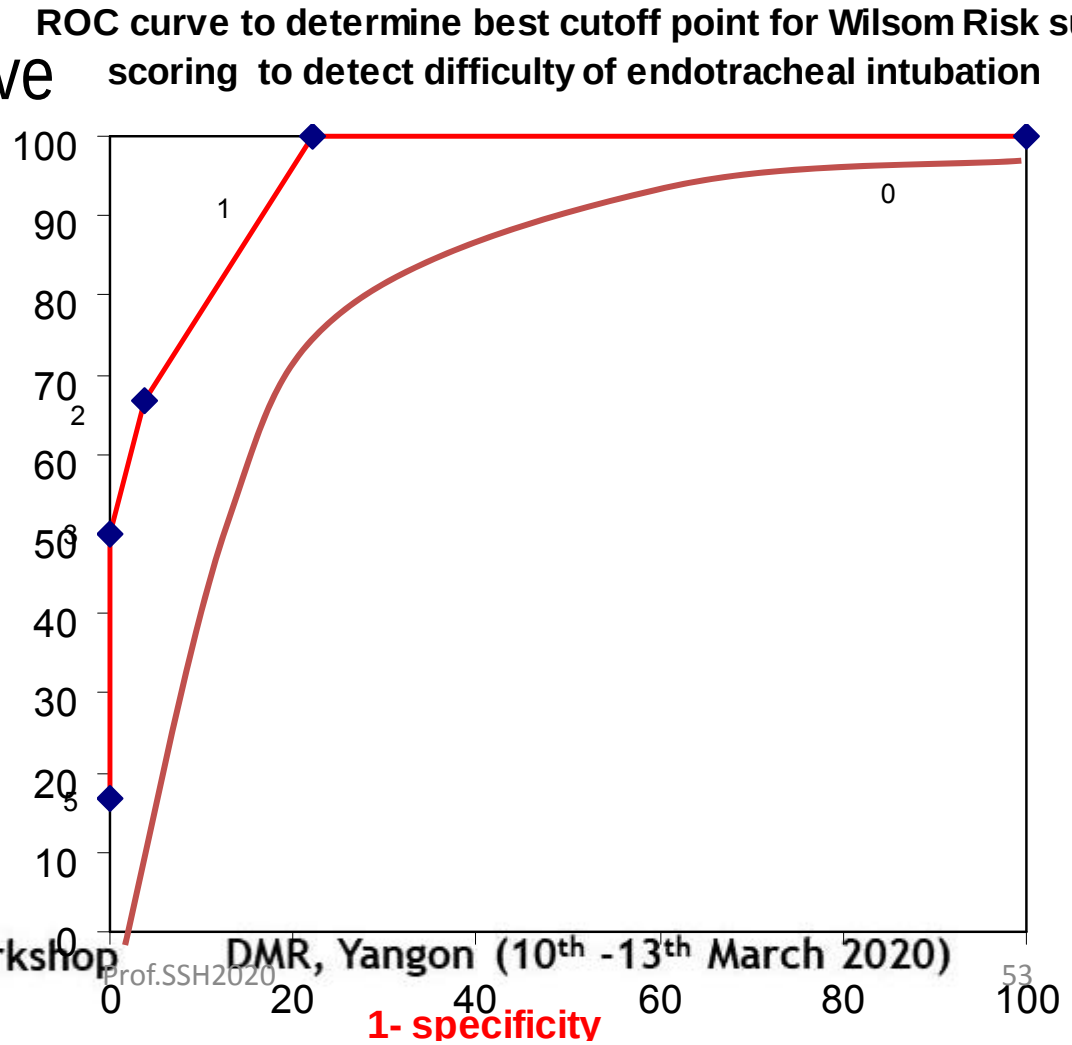
Receiver Operator Characteristic (ROC) Curve

- Plot true positive rate (sensitivity) against false positive rate (1-specificity) for several choice of positively criterion
- choose closest to top left corner to maximized the discriminative ability of the test



Receiver Operator Characteristic (ROC) Curve

- The area under the curve represent overall accuracy of the test
- useful to compare two test



Likelihood Ratios

- The likelihood that a given test result would be expected in a patient with a target diseases compare with the patient without a target diseases
- LR^+ : How much more likely is a positive test to be found in a person with the condition than in a person without it?
- Unlike predictive values, likelihood ratios are not influenced by prevalence of the disease.

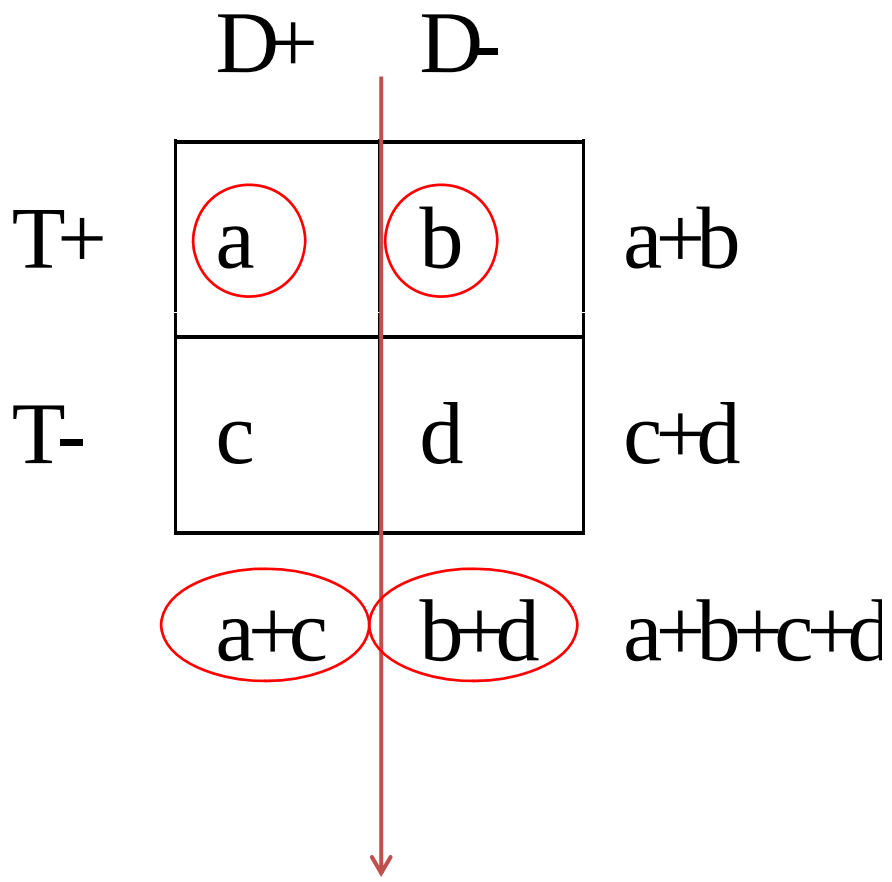
Likelihood Ratios Positive

Likelihood ratio positive (LR+) is the ratio of the sensitivity of a test to the false positive error rate of the test

LR+ = [a/(a+c)] / [b/(b+d)]

LR+ is the ratio of something that the clinicians do want to something that the clinicians do not want .

The higher the ratio is the better the test.



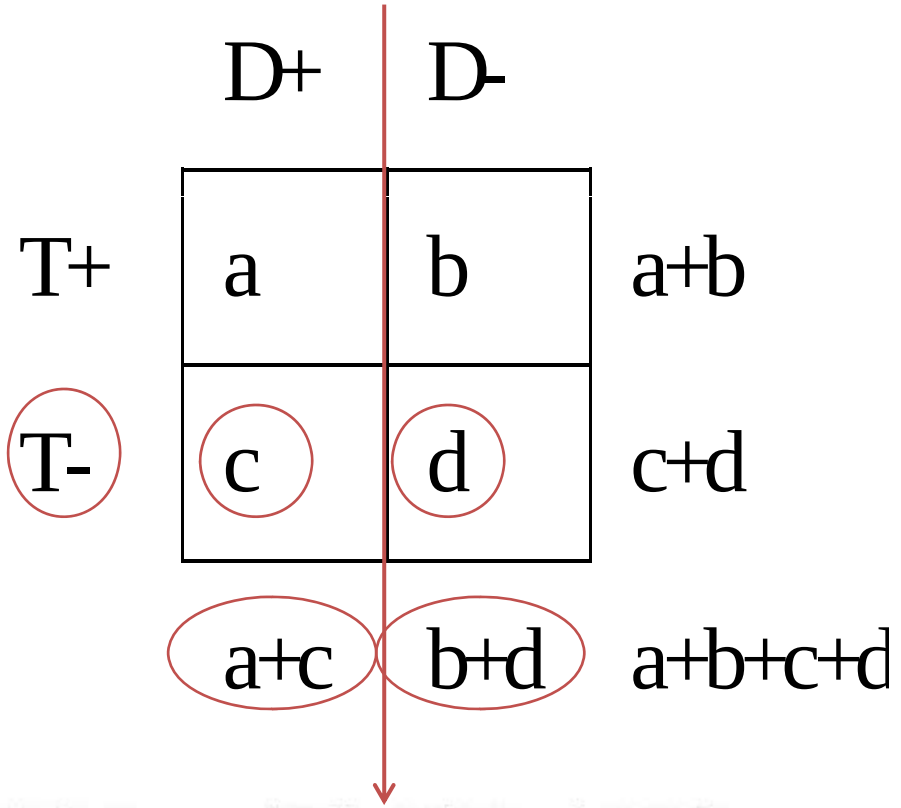
Likelihood Ratios Negative

Likelihood ratio negative (LR-) is the ratio of the false negative error rate of a test to the specificity of the test

LR- = [c/(a+c)] / [d/(b+d)]

LR- is the ratio of something that the clinicians do not want (false negative error) to something that the clinicians do want (specificity)

The closer the ratio is to 0 the better the test.



Measuring agreement

		Observer 1		
		Pos	Neg	Total
Obs 2	Pos	a	b	a + b
	Neg	c	d	c + d
		a + c	b + d	a+b+c+d

Observed agreement = a+d =Ao

Max. possible agreement = a+b+c+d = N

Overall percent agreement = (a+d) / (a+b+c+d)

Measuring agreement (Kappa)

		Observer 1		
		Pos	Neg	Total
Obs 2	Pos	a	b	a + b
	Neg	c	d	c + d
		a + c	b + d	a+b+c+d

cell a agreement expected by chance = (a+b)*(a+c) / (a+b+c+d)

cell d agreement expected by chance = (c+d)*(b+d) / (a+b+c+d)

Ac: Total agreement expected by chance = cell a agreement expected .. + cell d agreement expected ...

(Ao - Ac) / (N- Ac) = Kappa

Interpreting Kappa statistics

The kappa statistics measurement of agreement is scaled to be:

0 when the amount of agreement is what would be expected to be observed by chance, and,

1 when there is perfect agreement

Size of κ	*agreement*
0.81 to 1.00	**Almost perfect**
0.61 to 0.80	**Substantial**
0.41 to 0.60	**Fair**
0.21 to 0.40	**Slight**
0.00 to 0.20	**Poor**

**** Suggested by Landis and Koch (1977a, 165)***

Five Approaches to Expressing Prognosis

- Case-fatality
- 5-year survival
- Observed survival
- Median survival time
- Relative survival

What is survival analysis?

Statistical methods for analyzing longitudinal data on the occurrence of events (death, injury, onset of illness, recovery from illness)

data from randomized clinical trial or cohort study design.

- Kaplan-Meier curve
- log-rank test
- Hazard Ratios/Cox regression