LIES, DAMN LIES, AND STATISTICS

Kenneth Alonso, MD, FACP

How to lie with statistics

- Is public education failing as measured by the SAT?
- The SAT has been given for years.
- The range of possible scores is 200-800.
- The mean on the validated population is 500.
- The validated population are “advantaged” students.
- On a baseline exam given before intervening to assist disadvantaged students, the mean score is 490.
- On the subsequent exam, the mean score is 490.
- Has the intervention failed? Has the public educational system failed?

How to lie with statistics

- Customarily, advantaged students take the exam.
- Assume 10,000 take the exam; the mean score, as expected, is 500. Assume 1,000 disadvantaged also take the exam; their mean score is 400.
- The mean for this exam is 490. ([500x10,000 + 400x1,000]/11,000)

How to lie with statistics

- Following efforts to assist disadvantaged students, 2,000 take the next exam. Their mean score is 450.
- 12,000 advantaged students also take the same exam. Their mean score is 500, as expected.
- The mean for this exam is 490. ([500x12,000+450x2,000]/14,000)
- SAT means have not changed.
- However, the gap between populations has narrowed significantly.
- The intervention has been successful.
- The public educational system has not failed.

How to lie with statistics

- Equal pay legislation obligates employers to pay men and women at the same rate for the same job.
- Women as a group earn 80% of what men earn.
- However, young single women earn more than their male counterparts.
- Prior to the last two decades, fewer women than men went to college.
- Now the situation has reversed following social intervention.
- What should be the next public policy step?

How to evaluate a medical publication

Kenneth Alonso, MD, FACP

How to evaluate an article about a diagnostic test

- Has there been an independent (“blind”) comparison a criterion standard of diagnosis?
- Has the diagnostic test been evaluated in a patient sample that includes a appropriate spectrum (prevalence) of mild and severe, treated and untreated disease, as well as individuals with different but commonly confused disorders?
- Was the study setting as well as the filter through which the patients passed adequately described?
- Has the reproducibility of the test result (precision) and its interpretation (observer variation) been determined?

How to evaluate an article about a diagnostic test

- Has the term “normal” been defined sensibly as it applies to this test?
- If the test is advocated as part of a cluster or sequence of tests, had its individual contribution to the overall validity of the cluster and sequence been determined?
- Have the tactics for carrying out the test been described in sufficient detail to permit their exact replication?
- Has the utility of the test been determined?

- Simel, DL, Drummond,R, The rational clinical examination. Evidence based clinical diagnosis. McGraw-Hill (New York) for the American Medical Association (Chicago). 2009. p3

Observer variation

- For an item to be accurate, it must be reproducible (precision).
- Is it repeatable (intra-observer) or do two or more observers agree on the presence or absence of symptom or sign (inter-observer).
- How do we evaluate the reliability of agreement?
- How do we determine agreement is not by chance?
- Utilizing a 2x2 table, agreement between two observers would be, for true positives, [(a+b)x(a+c)]/(a+b+c+d)
- For true negatives, [(c+d)x(b+d)]/(a+b+c+d)

Observer variation

- The expected agreement is true positive agreement and true negative agreement divided by the number studied
- {[(a+b)x(a+c)]/(a+b+c+d) + [(c+d)x(b+d)] / (a+b+c+d)}/(a+b+c+d)
- Agreement beyond chance, κ, is:

  (observed agreement – expected agreement) / (1 – expected agreement)
- The higher the level of κ, the better the agreement.

Observer variation

- A value of -1.0 indicates complete disagreement, while a value of 1.0 indicates complete agreement; a value of 0.0 is chance.
- A κ of >0.6 indicates substantial agreement, reproducibility.
- Precision, however, does not mean accuracy.

EVALUATION OF MEDICAL TESTS AND TREATMENT

REFERENCE RANGES

Kenneth Alonso, MD, FACP

Normal

- What constitutes a "normal" group?
- Hospitalized patients?
- Ambulatory patients?
- Patients processed through an ambulatory clinic?
- Medical students?
- Members of the armed services?
- The selection of the "normal" or reference group is critical to the power of the test to discriminate disease and non-disease states.

Reference range

- Measurements are made on a population. The distribution of those measurements reflect the variance of the population sampled.
- The MEAN [x] and the STANDARD DEVIATION [s] about that MEAN are the two parameters that characterize the population distribution.
- The STANDARD DEVIATION OF THE MEAN [sem] is calculated by dividing the standard deviation by the square root of the number sampled.
- It reflects the variance of the measurement, not the individual.
- It diminishes as the number examined rises.

Reference range

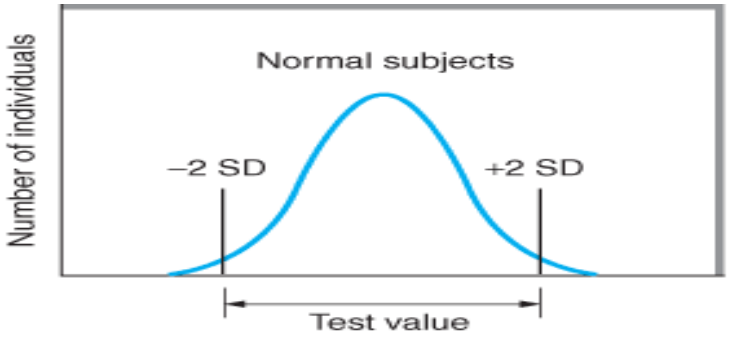
- Measurements made on a population generally follow a “normal” or Gaussian [bell curve] distribution.
- Measurements that are skewed in one direction may be first "normalized" by taking their log or sine values.
- Then the MEAN and STANDARD DEVIATION can be calculated.
- This avoids setting a low end of the range below zero.
- Alternatively, a skewed sample may be characterized by its MEDIAN as well as its values at 25^{th} and 75^{th} percentiles.

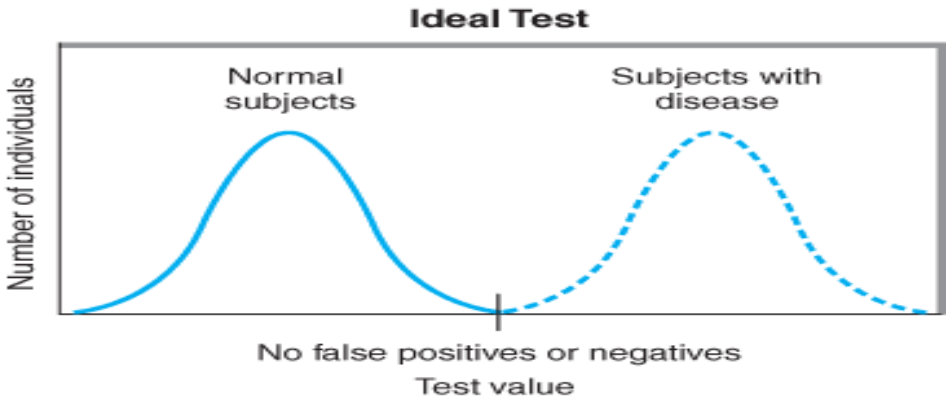
Reference range caveats

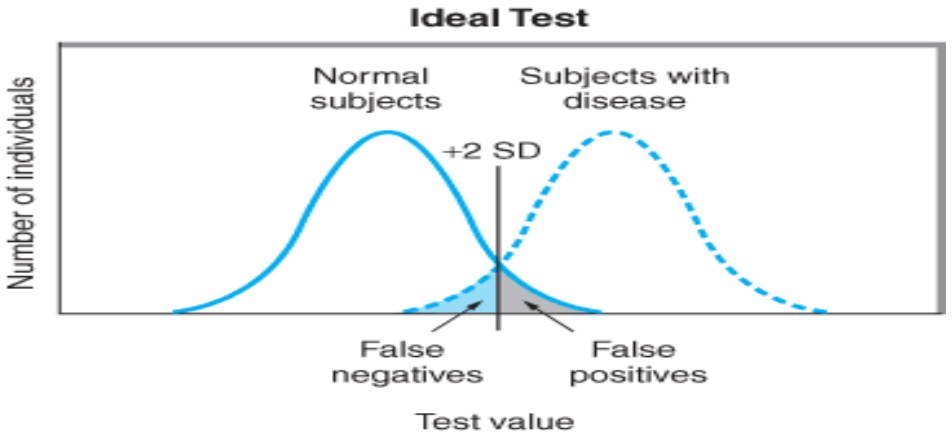
- Published laboratory normal ranges are determined principally from young healthy white men in their 20’s (on serum).
- Levels obtained on the same person vary with the time of day the specimen is obtained.

Reference range caveats

- Numbers of lymphocytes vary with age (higher in children).
- Uric acid, ALT levels do vary with sex (lower in women).
- Bleeding times vary with altitude (longer at higher altitudes).
- Aldosterone and free testosterone levels decrease with age.







Source: Gardner DG, Shoback D: *Greenspan's Basic and Clinical Endocrinology*, 8th Edition: http://www.accessmedicine.com

Fig. 4-1 Accessed 08/01/2010

Receiver operating curve

- This is a graphical method accounting for the mutual dependence between sensitivity and specificity.
- It evaluates the extent to which variation in sensitivity and specificity can be explained by variation in positivity thresholds.
- Laboratory values are compared to confirmed diseased and non-diseased states.
- Those values which separate the largest number of patients with disease from those in whom the disease is absent are selected as the action limits.

Receiver operating curve

- Alternatively, an action level may be set separating those with and without disease.
- The area under the curve (AUC) as compared to that of a curve whose results reflect the fact that the true-positive and false-positive results are the same is used as a measure of the diagnostic performance of the test
- A 45° diagonal line through 0 is indicative of a result through chance alone
- The greater the AUC, the better the test.

Receiver operating curve

- In an ROC curve, the true positive rate (sensitivity) is plotted on the vertical axis, and the false-positive rate (1 – specificity) is plotted on the horizontal axis for different cutoff points for the test.
- The closer an ROC curve is to the upper left-hand corner of the graph, the more accurate it is, because the true-positive rate is 1 and the false-positive rate is 0.

Receiver operating curve

As the criterion for a positive test becomes more stringent, the point on the curve corresponding to sensitivity and specificity (point A) moves down and to the left (lower sensitivity, higher specificity); if less evidence is required for a positive test, the point on the curve corresponding to sensitivity and specificity (point B) moves up and to the right (higher sensitivity, lower specificity).

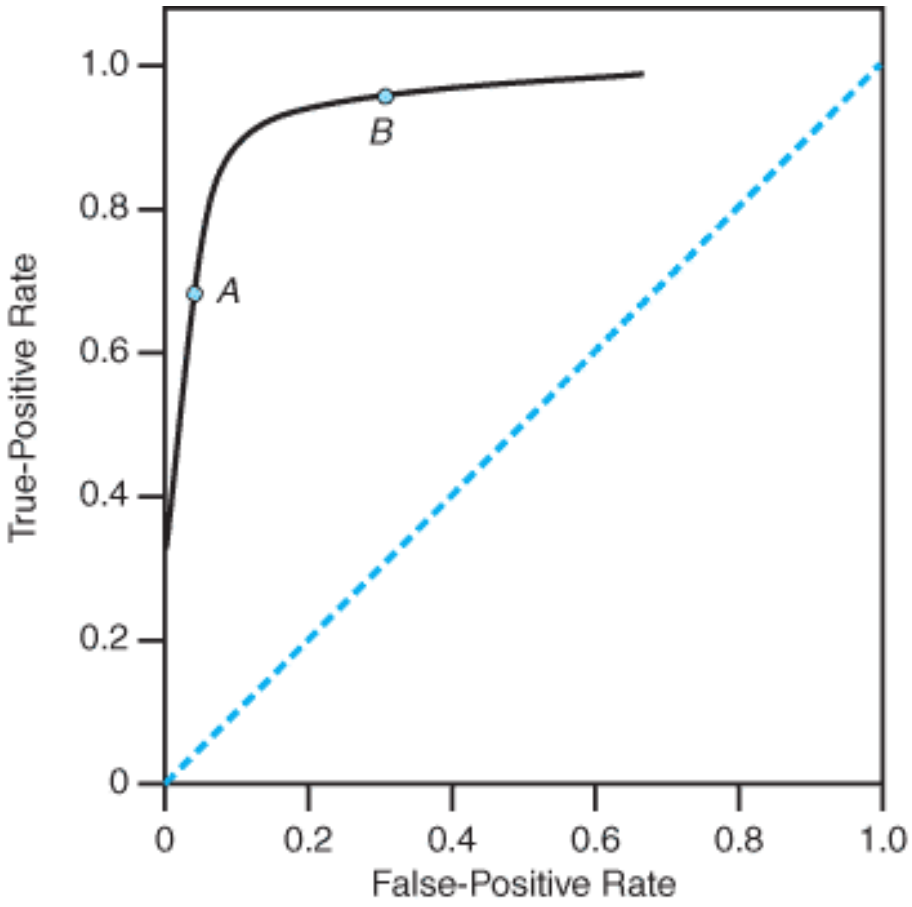
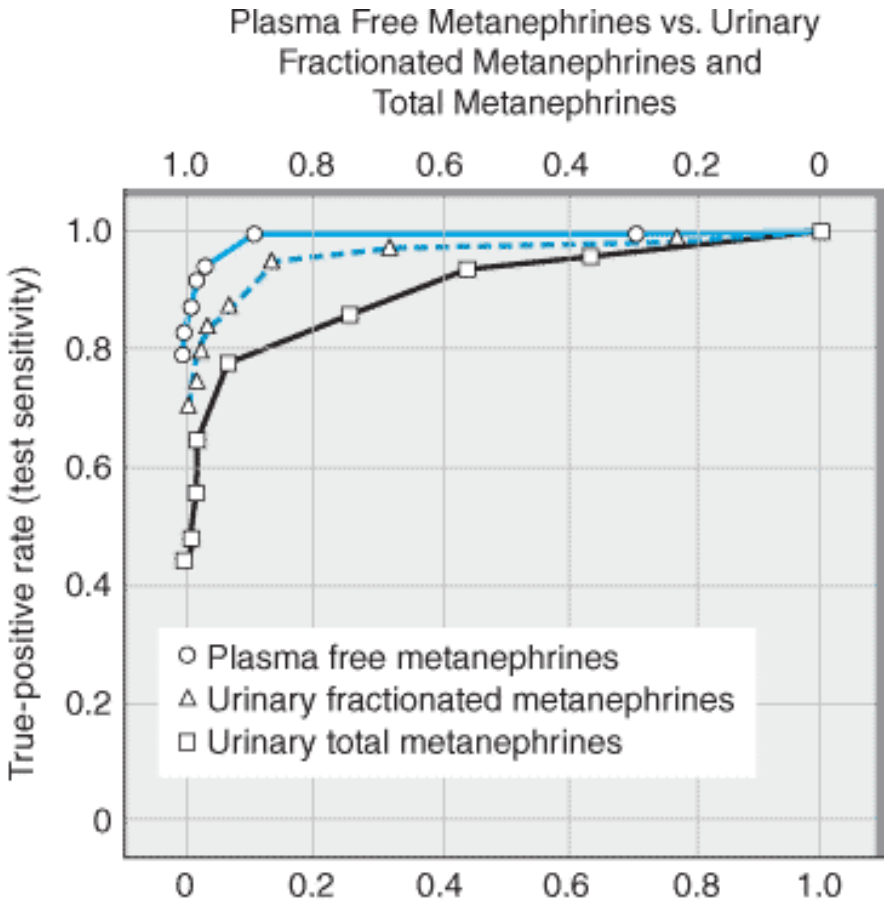


Fig. 4-5 Accessed 08/01/2010

Receiver operating curve



Source: Gardner DG, Shoback D: *Greenspan's Basic and Clinical Endocrinology*, 8th Edition: http://www.accessmedicine.com

In this comparison of the performance of different tests for the diagnosis of pheochromocytoma, plasma free metanephrines performed better at any given cutoff point than urinary fractionated metanephrines or urinary total metanephrines.

(Modified with permission from Lenders JWM et al: Biochemical diagnosis of pheochromocytoma: which test is best? JAMA 2002;287:1427.)

Fig. 4-6 Accessed 08/01/2010

EVALUATION OF MEDICAL TESTS AND TREATMENT

PREDICTIVE VALUES

Kenneth Alonso, MD, FACP

Clinical Professor, Morehouse School of Medicine, Atlanta, Georgia

Clinical Professor, LECOM Bradenton, Florida

Prevalence

- The PREDICTIVE VALUE of a test, not its SENSITIVITY or SPECIFICITY, is the clue to its utility.
- The value of a test is dependent upon the population studied.
- The PREVALENCE of a disease is critical to the utility of laboratory tests and the use of medical resources.
- PREVALENCE is the percentage of patients who have the target disorder. (Total cases/ Total population at risk.) This is PRE-TEST PROBABILITY.

Bayes 2x2 table

		Disease	
		+ Present	– Present
Test result	+	a True positive (TP)	b False positive (FP)
	–	c False negative (FN)	d True negative (TN)

Sensitivity = a/(a + c) = TP/(TP + FN)

Specificity = d/(b + d) = TN/(TN + FP)

Positive predictive value = a/(a + b) = (TP/TP + FP)

Negative predictive value = d/(c + d) = TN/(TN + FN)

Diagnostic accuracy = (a + d)/(a + b + c + d) = (TP+ FN)/(TP + FP + TN + FN)

Pre-test (prior) Probability = (a + c)/(a + b + c + d) = (TP + FN)/(TP + FP + TN + FN)

Likelihood ratio for a positive test result (LR+):

= [a/(a + c)]/[b/b(b + d)] = TP rate/FP rate = [TP/(TP + FN)/(FP/TN + FP)] =

= sensitivity/(1 – specificity)

Likelihood ratio for a negative test result (LR–):

= [c/(a + c)]/[d/(b + d)] = FN rate/TN rate = [FN/(TP + FN)]/[TN + FP)] =

= (1 – sensitivity)/specificity

Source: Gardner DG, Shoback D: *Greenspan's Basic and Clinical Endocrinology*, 8th Edition: http://www.accessmedicine.com

Fig. 4-2 Accessed 08/01/2010

Sensitivity

- SENSITIVITY is the percentage of patients WITH the target disorder who have a POSITIVE test result.
- SENSITIVITY is calculated as the number of known patients with disease who have a positive test result, divided by the total number of known patients with disease in the population sampled.
- a/a+c

Specificity

- SPECIFICITY is the percentage of patients WITHOUT the target disorder who have a NEGATIVE test result.
- SPECIFICITY is calculated as the number of known patients without disease who have a negative test result, divided by the total number of known patients without disease in the population sampled.
- d/b+d

Sensitivity and specificity

- A test that is highly SENSITIVE identifies (virtually) all patients WITH disease.
- A test that is highly SPECIFIC identifies (virtually) all patients WITHOUT disease.

Sensitivity and specificity

- A test that is both highly sensitive and highly specific may not be of clinical utility, however, if the prevalence of disease in the population studied is very, very low.
- In such a case, a POSITIVE result does not predict the presence of disease.
- In such a case, a NEGATIVE result is highly predictive of the absence of disease. The number of false positives is high, however; there are few false negatives.

Choice of control groups

		Cushing's Syndrome	
		Present	Absent
1mg overnight dexamethasone suppression	No suppression	151	5
	Suppression	3	461

Sensitivity = 151/(151 + 3) = 98.1%

Specificity = 461/(5 + 461) = 98.9%

A

		Cushing's Syndrome	
		Present	Absent
1mg overnight dexamethasone suppression	No suppression	151	101
	Suppression	3	858

Specificity = 858/(101 + 858) = 89.5%

B

		Cushing's Syndrome	
		Present	Absent
1mg overnight dexamethasone suppression	No suppression	151	96
	Suppression	3	397

Specificity = 397/(96 + 397) = 80.5%

C

Source: Gardner DG, Shoback D: *Greenspan's Basic and Clinical Endocrinology*, 8th Edition: http://www.accessmedicine.com

Diagnosis of Cushing's syndrome with the 1-mg overnight dexamethasone suppression test: test characteristics with normal controls (Panel A); all controls (Panel B); and "obese" and "other" controls (Panel C). These data show how the specificity of the test is affected by the types of control subjects.

(Reproduced with permission from Crapo L: Cushing's syndrome: a review of diagnostic tests. Metabolism 1979;28:955.)

Fig. 4-3 Accessed 08/01/2010

Sensitivity and specificity

- Many screening studies in the US have shown the prevalence of HIV infection (seropositive) to be 0.4%.
- The rapid screening test is 99.6% sensitive and 97.5% specific for this condition.

- Consider the results of HIV screening of ten thousand people selected at random.

HIV

Number of patients	Test positive	Test negative
Infected	0038	0002
Not Infected	0388	9572

Sensitivity and specificity

- SENSITIVITY

  38/40 patients classified correctly.

- SPECIFICITY

  9572/9960 patients classified correctly.

- Ten thousand patients examined.

Predictive value

- The POSITIVE PREDICTIVE VALUE is calculated as a/a+b.
- The NEGATIVE PREDICTIVE VALUE is calculated as d/c+d.

- The FALSE POSITIVE RATE is calculated as b/b+d. These are those patients who do not have the disease but who have tested positive.
- The FALSE NEGATIVE RATE is calculated as c/a+c. These are those patients who have the disease but who have tested negative.

Predictive value

- POSITIVE PREDICTIVE VALUE

  426 patients classified as HIV positive; however, only 38 are infected with HIV.

  Thus, 380 patients are classified incorrectly, the false positives.

- NEGATIVE PREDICTIVE VALUE

  9572/9574 patients classified as HIV negative

  Two patients are classified incorrectly, the false negatives.

  This may be acceptable for screening blood donors as any questionable units will be discarded.

  It is not a good strategy to decide upon quarantine.

When prevalence changes

- If the prevalence of HIV is 1%, then testing will uncover 99/100 infected patients (one will still not be uncovered).
- 99 patients not infected will be categorized inappropriately; 9801 will truly be negative.
- The positive predictive value has risen to 50%; the negative predictive value is 99%.
- INCIDENCE refers to new cases only as those previously identified are no longer considered at risk.
- Differs from PREVALENCE.

When prevalence changes

Number of patients	Test positive	Test negative
Infected	0099	0001
Not Infected	0099	9801

Predictive value

Disease prevalence or pretest probability	0.1%	1%	10%	50%	90%
Positive predictive value	0.89%	8.33%	50.0%	90.0%	98.78%
Negative predictive value	99.99%	99.89%	98.78%	90%	50.0%

Source: Gardner DG, Shoback D: *Greenspan's Basic and Clinical Endocrinology*, 8th Edition: http://www.accessmedicine.com

Positive and negative predictive values as a function of disease prevalence, assuming test sensitivity and specificity of 90% for each.

Fig. 4-7 Accessed 08/01/2010

HIV screening recommendations

- Screen all newborns of HIV positive mothers. (PCR)
- Screen in those populations where prevalence of HIV positive is >1% (sexually transmitted disease clinics, prisons).
- p24 antigen testing is the screen of choice if acute illness suspected. Do not use viral load tests.
- Screen all pregnant women.

HIV screening caveats

- Recent CDC recommendations for universal screening assume the rapid screening antibody test cost is very low (as must be the confirmatory Western Blot)
- AND there is no negative impact from a false positive result
- AND that risk behavior is changed by the result
- AND anti-retroviral therapy completely suppresses HIV in body fluids and thus limits infectivity.
- Those assumptions are not supported by clinical data.

EVALUATION OF MEDICAL TESTS AND TREATMENT

PROBABILITIES

Kenneth Alonso, MD, FACP

Probability of more than one event occurring

- The probability that an individual has one OR another mutually exclusive property is the sum of the probabilities of having each individual property. (Man or woman?)
- The probability that an individual has one OR another property not mutually exclusive is the sum of the probabilities of having each individual property minus the probability of having both properties. (Brown eyes or blue eyes or one of each?)

Probability of more than one event occurring

- The probability that an individual has one AND another property is the product of the individual probabilities. (Smart and pretty?)
- For events whose occurrence is dependent upon the occurrence of an earlier event, probabilities are multiplied together. (Graduated from high school, college, admitted into medical school.)

Likelihood ratio

- The LIKELIHOOD RATIO is the ratio of the probability of a test result among patients with the target disorder to the probability of that same test result among patients who are free of the target disorder.
- For a POSITIVE result, the Likelihood Ratio is calculated as: [sensitivity/ (1 – specificity)].

  (a/a+d)/(c/c+d)
- For a NEGATIVE result, the Likelihood Ratio is calculated as: [(1 – sensitivity)/ specificity].

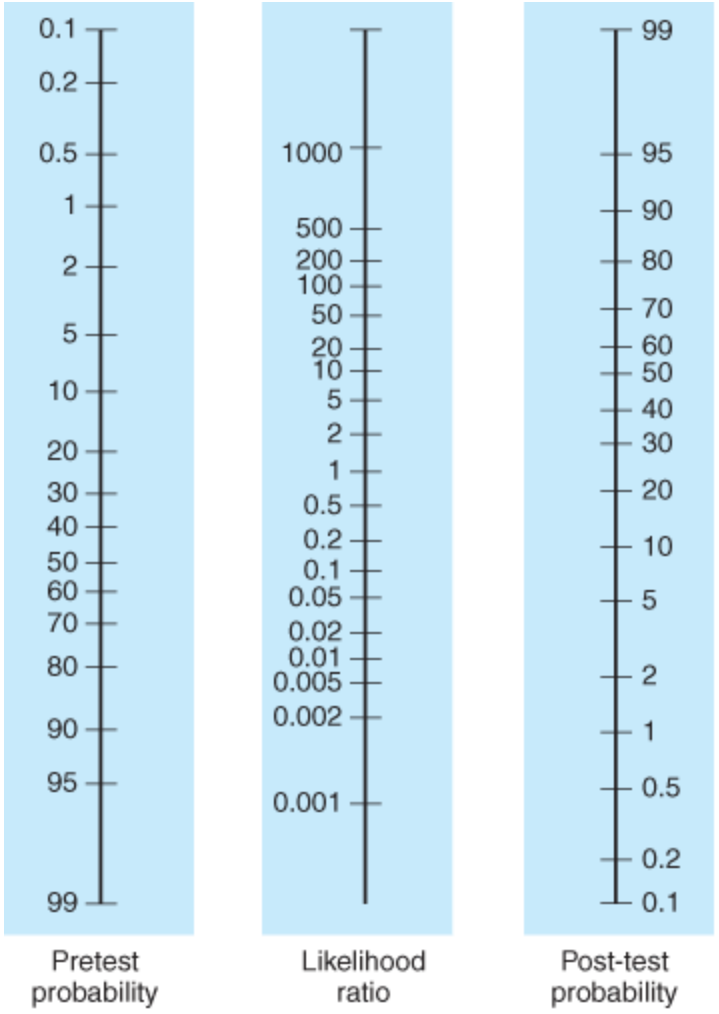
  (b/a+d)/(d/c+d)

Likelihood ratio

- A high likelihood ratio may not be clinically useful.
- Consider a study reporting the use of D-dimer to distinguish those symptomatic patients with pulmonary embolism from those without pulmonary embolism.
- 2311 symptomatic patients were examined.
- 118 of those patients had confirmed pulmonary embolism.
- The prevalence of disease is 5.1% in this symptomatic group.
- This is 20 times higher than in an asymptomatic population (likelihood ratio >20).

Likelihood ratio

- 1 of 2193 symptomatic patients without pulmonary embolism had a positive D-dimer assay.
- The negative predictive value of a negative assay is 99%.
- However, only 2 of 118 symptomatic patients with pulmonary embolism were identified by a positive D-dimer assay alone.
- Despite a high likelihood, a positive D-dimer result adds little to the diagnostic strategy.
- A negative result, however, excludes pulmonary embolism.



Source: Gardner DG, Shoback D: *Greenspan's Basic and Clinical Endocrinology*, 8th Edition: http://www.accessmedicine.com

Nomogram for likelihood ratios.

(Adapted from Fagan TJ: Nomogram for Bayes theorem. New Engl J Med 1975;293;257. Reprinted, with permission of The New England Journal of Medicine. Copyright 1975, Massachusetts Medical Society.)

Fig. 4-7 Accessed 08/01/2010

Likelihood ratio

- The pre-test probability of malignancy in a breast lump in a 25 year old woman without a family history of breast or ovarian cancer is 1%.
- The sensitivity of mammography is 70%.
- The specificity of mammography is 40%.
- The post-test likelihood of finding malignancy with mammography alone in a 25yo woman with no family history of breast or ovarian cancer is:

  1/100 x 40/(1-0.7)= 1.3/100 or 1.3%.

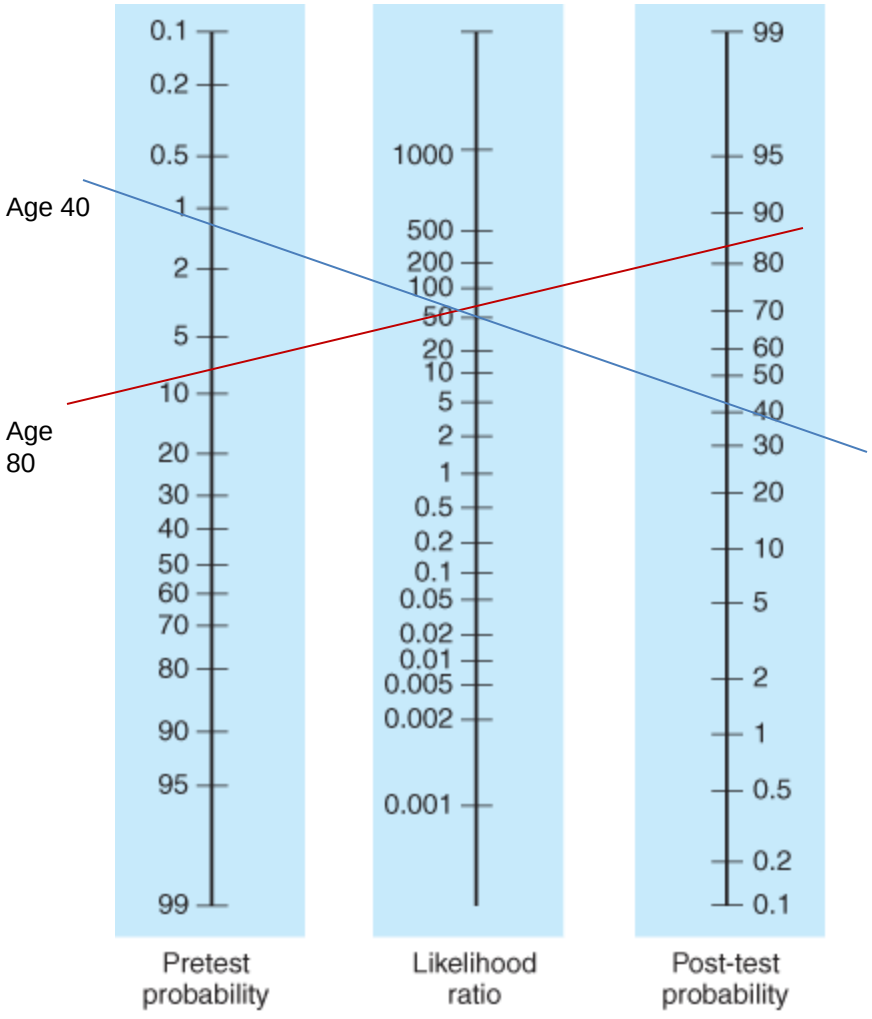
Predictive value

Disease prevalence or pretest probability	0.1%	1%	10%	50%	90%
Positive predictive value	0.89%	8.33%	50.0%	90.0%	98.78%
Negative predictive value	99.99%	99.89%	98.78%	90%	50.0%

Source: Gardner DG, Shoback D: *Greenspan's Basic and Clinical Endocrinology*, 8th Edition: http://www.accessmedicine.com

Positive and negative predictive values as a function of disease prevalence, assuming test sensitivity and specificity of 90% for each.

Fig. 4-7 Accessed 08/01/2010



Source: Gardner DG, Shoback D: *Greenspan's Basic and Clinical Endocrinology*, 8th Edition: http://www.accessmedicine.com

Nomogram for likelihood ratios. Breast cancer.

(Adapted from Fagan TJ: Nomogram for Bayes theorem. New Engl J Med 1975;293;257. Reprinted, with permission of The New England Journal of Medicine. Copyright 1975, Massachusetts Medical Society.)

Fig. 4-7 Accessed 08/01/2010

Modified

When prevalence changes

- BY THE AGE OF 90
- One of every two men will have suffered from prostate cancer.
- One of every eight women will have suffered from breast cancer.

- BELOW THE AGE OF 50
- The prevalence of either cancer is 1%.

When prevalence changes

- SCREENING AFTER THE AGE OF 50
- For the entire population of women this translates into a gain of 8 days of life.
- For the affected woman, this translates into a gain of 8 years of life.
- The cost of finding a new breast cancer in women over the age of 50 is $50,000.

When prevalence changes

- As it has been demonstrated that repeat screening every three years is as effective as yearly screening, the cost of finding a new cancer should be less.
- The rate of false positives rises to as high as 16% in populations rescreened several times over the years.
- Screening is not suggested for those whose life expectancy is less than 10 years.

When prevalence changes

- SCREENING AFTER THE AGE OF 50
- It has not been demonstrated that the use of PSA screening for prostate cancer is beneficial.
- PSA screening for prostate cancer is not recommended for men over the age of 65 if the initial PSA is less than or equal to 1 ng/ml.
- Screening is not recommended if the life expectancy of the patient is less than 10 years.
- The cost of finding a new prostate cancer is $35,000.

When prevalence changes

- More women will die of heart disease than will die of breast cancer.

- More men will die of heart disease than will die of prostate cancer.

Risk ratio

- Relative Risk is the ratio between the rate of the outcome in the treated group and the rate of the outcome in the control group.
- For adverse outcomes, a ratio <1.0 favors the treatment group.
- This is calculated as (a/a+b) / (c/c+d).
- Attributable risk is the difference in risk between exposed and unexposed populations (or the proportion of disease occurrences as a result of exposure).

Odds ratio

- The Odds ratio is the ratio of the odds of the outcome in a treated (or exposed) group and the odds in the control group.
- This is calculated as a x d / b x c.
- Odds ratio always overestimates relative risk.
- As the baseline probability increases and the relative risk increases, divergence is marked.

Number needed to treat or harm

- Absolute risk difference is the difference between post-exposure and baseline risk.
- Number needed to treat or harm is the reciprocal of the absolute risk difference.
- If the absolute risk is 0.10, then 10 patients exposed to treatment would yield benefit or harm to 1 patient.

Number needed to treat or harm

- To calculate the number needed to treat from the relative risk requires the baseline incidence of the complication.
- If the adverse event occurs 1% of the time
- And there is a Relative Risk reduction of 50% (the absolute risk reduction is 0.5%)
- Then the Number Needed to Treat is 1/.005 or 200 patients to see a benefit.

Number needed to treat or harm (cost)

- To calculate the cost of an intervention, one must have the number needed to treat, the length of time needed to treat in order to see a benefit, and the cost of the intervention.
- If 200 is the number needed to treat, 3 years is the length of time needed to treat in order to see a benefit, and the cost of the intervention is $1/day, then

  200 x 3 x 365 = $21,900

EVALUATION OF MEDICAL TESTS AND TREATMENT

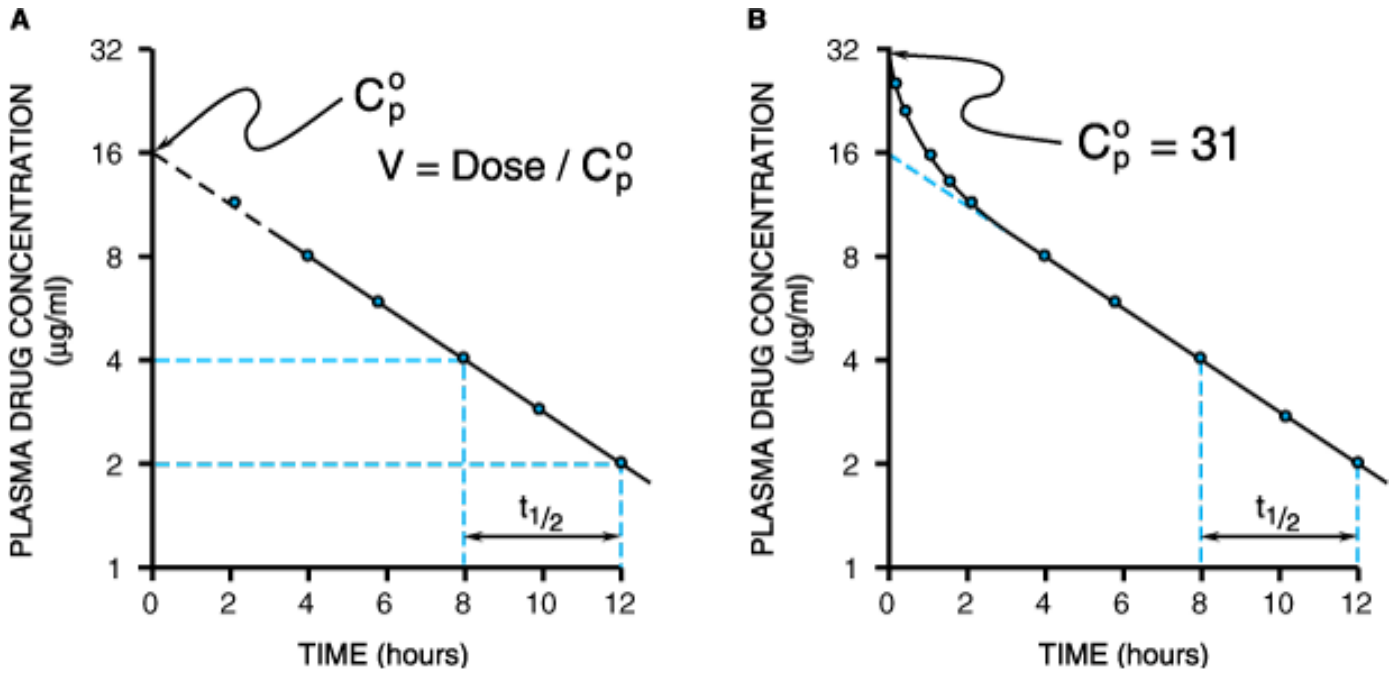
ERRORS

Kenneth Alonso, MD, FACP

Clinical Professor, Morehouse School of Medicine, Atlanta, Georgia

Clinical Professor, LECOM Bradenton, Florida

Sampling error



Because of late sampling times, multicompartment distribution has been ignored.

Fig. 1-3 Accessed 08/01/2010

Statistical errors

- Type I error
- A Type I error is an error in the true sense. A conclusion is drawn that the null hypothesis is false when, in fact, it is true.
- Type II Error (β)
- A Type II (β) error is a potential failure to reject a false null hypothesis.

Confidence interval

- The more an experimenter protects him or herself against Type I errors by choosing a low level, the greater the chance of a Type II error.
- Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected.
- However, it increases the chance that a false null hypothesis will not be rejected, thus lowering power.

Confidence interval

- The Type I error rate is almost always set at a p-value of 0.05 or 0.01.
- The latter is more conservative as it requires stronger evidence to reject the null hypothesis at the 0.01 level.
- A properly structured study asks a question that requires a yes/no answer (the null hypothesis).
- The assumption is that the null hypothesis is true.
- Studies are structured to avoid errors in rejecting or accepting the null hypothesis.

Confidence interval

- A p-value is frequently reported.
- This represents the confidence interval (mean ± standard deviation) that the result obtained is not by chance.
- A p-value of 0.05 (two standard deviations about the mean) demonstrates that a result outside the confidence interval is not due to chance is at a probability level of 5%.
- A p-value >0.05 means that the null hypothesis is statistically consistent with the observed result.

Confidence interval

- If the investigator has set a level of significance of significance of 0.01 (three standard deviations about the mean) and reports a p-value of 0.02, the investigator has rejected the null hypothesis at that level
- The rejection is certain.
- Conversely, had a level of significance of 0.05 been set, the investigator would have accepted the null hypothesis at that level; the acceptance is certain.
- The p-value is related to the level of error one is willing to tolerate.

Confidence interval

- If the 95% confidence interval for a mean difference between two variables includes 0, then there is no significant difference noted.
- If the 95% confidence interval for odds ratio or relative risk includes 1, there is a significant difference noted.
- There is no significant difference if the values overlap.
- As correlation coefficients approach 1, the greater the correlation.

Biases

- INTENT TO TREAT
- Surgery is proposed for a condition.
- The endpoint is time from diagnosis to death.
- Patients who die before the surgery are included in the no surgery group.
- Survival curves will favor the surgical group whether surgery is effective or not.
- LEAD TIME BIAS
- A test detects a disease earlier than current methods.
- Earlier intervention does not change the course of the disease.
- If one were to examine survival from time of diagnosis, it would appear that the test is helpful.

EVALUATION OF MEDICAL TESTS AND TREATMENT

COMPARISONS

Kenneth Alonso, MD, FACP

When do serial values differ?

- The COEFFICIENT OF VARIATION is the standard deviation of the sample divided by the mean.
- A coefficient of variation (CV) of 10% is common in laboratory testing.)
- Electrolyte determinations employing ion specific electrodes have a CV of 1%.
- Automated cell counts and cell size determinations also have a CV of 1%.
- Thus, a difference in serial values greater than 2CV is generally regarded as significant.

- <u>Laboratory values are not absolute numbers.</u>

Comparison of groups

- Similar experiments, with similar null and alternative hypotheses, will be analyzed differently depending upon the property examined.
- If the property can be measured, it should be analyzed with a t-test or with an ANOVA.
- If the property is an attribute or a category, it should be analyzed with a Chi-square test. For a 2 x 2 contingency table, the formula follows:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Comparison of groups

- For a group of n=1, the paired t-test is chosen.
- To compare 2 groups, the unpaired t-test is often chosen.
- If 3 or more groups are to be compared, ANOVA is the procedure of choice.
- If multiple comparisons are to be made, ANOVA is repeated with each comparison.

Wilcoxson t-test

- Normal data distribution is assumed.
- Samples may be Independent (two randomly selected unrelated groups), or
- Dependent (two groups matched for some variable or repeated measurements on the same group).
- The degrees of freedom for the test are 2(n-1) where n is the sum of the number of participants. SDp is the pooled standard deviation.

$$t_{(n_1+n_1-2)} = \frac{(\bar{X}_1 - \bar{X}_2)}{SD_p\sqrt{[(1/n_1)+(1/n_2)]}}$$

$$SD_p = \sqrt{\frac{(n_1-1)\,SD_1^2+(n_2-1)\,SD_2^2}{n_1+n_2-2}}$$

ANOVA

- Normal distribution is assumed.
- Alternatively, each response can be ranked and ANOVA performed on rank-transformed data.
- This can reduce error in comparing samples that are not normally distributed.
- The degrees of freedom are 2n-1 where n is the sum of the number of participants.
- The F ratio is found by dividing the mean square among groups by the error mean square.

Pearson product moment correlation coefficient

- Indicates the strength and relationship of two random variables.
- Requires a normal distribution.
- A value of +1 means there is a perfect positive relationship between the two variables; -1, negative relationship; 0, no relationship.

$$r = \frac{\Sigma(X - \bar{X})\,(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2\,\Sigma(Y - \bar{Y})^2}}$$

If the distribution is not normal

- A CHI-SQUARE (χ^2) test is a better test to evaluate the strength of relationship between two variables if the distribution is not normal.
- The square of the correlation coefficient is known as the COEFFICIENT OF DETERMINATION, and is the fraction of the variance in y that is accounted for by a linear fit of x to y.

Size of study group

- Determine the p-value, probability, or Type I error rate (α).
- Determine the number of predictors.
- Determine the anticipated effect size, δ.
- By convention, 0.02, 0.15, and 0.35 are small, medium, and large, respectively.
- Determine the desired statistical power level. (Power, φ, is (1-Type II error rate, β)).
- By convention, this should be 0.80 or higher.
- The standard deviation between means, s, is roughly one-quarter of the mean difference.

Size of study group

- For one predictor the sample size needed, is:

  $n = (\Phi \times s / \delta \times \alpha)^2$

- A study with a yes/no outcome measure needs approximately 50 events to occur in the control group to have an 80% power of detecting a 50% relative risk reduction.
- If the control group risk is 20%, two groups of roughly 250 are required; if a 10% risk, 500; if a 5% risk, 1000; if a 1% risk, 5000.

Size of study group

- A study with a continuous outcome measure needs about 50 persons per group.
- The sample size required to detect minimum differences varies from 17 for 1 standard deviation; 33 for 0.7 standard deviation; 64 for 0.5 standard deviation; 175 for 0.3 standard deviation; 1571 for 0.1 standard deviation.

Is it useful?

- MEANS generally differ between groups and may be so as a result of chance.
- If confidence intervals do not overlap between groups, the two groups differ.
- Increasing the size of the group increases the likelihood that small differences will be detected.
- A twenty percent difference between groups can be demonstrated with a sample size of 50 patients.

Is it useful?

- The right question must be asked.
- Zidovudine was approved for AIDS treatment because of a difference of 11 AIDS defining events between the treated group of five hundred patients and that of the untreated group of similar size.
- Survival was not affected, not surprisingly, as mortality was not chosen as an end point.
- Quality of life diminished. That was not chosen as an endpoint.

HIV treatment results

- HAART treatment of HIV infection is very expensive. When does one initiate treatment?
- Early treatment of HIV infection with HAART is associated with increased life expectancy regardless of viral load if the patient is <30 years old and CD4 count is >200 cells/mm^3.
- Life expectancy ranges from 14.5 years if viral load >300,000 copies/ml
- (and rises) to 18.2 years if viral load <10,000 copies/ml and CD4 >500 cells/mm^3.

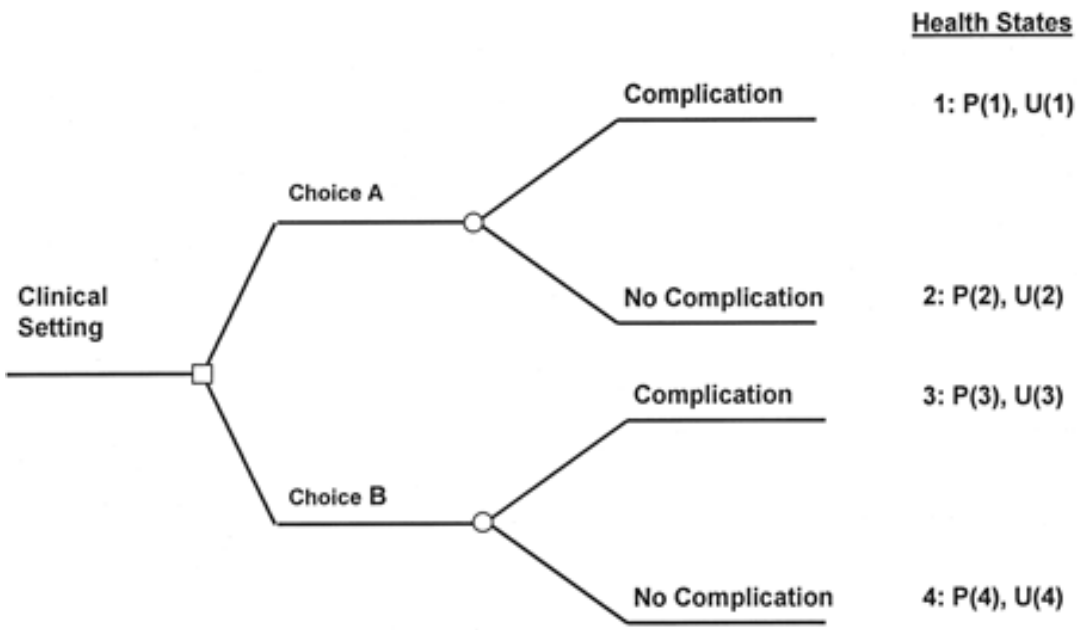
HIV treatment results

- Early treatment of HIV infection with HAART in patients OVER 40 YEARS OF AGE is associated with a life expectancy of 11.4 years if CD4 counts are >200cells/mm^3 AND viral loads are >300,000 copies/ml, rising to 12.9 years if CD4 counts are >500cells/mm^3 AND viral loads are <10,000 copies/ml.
- Little improvement in the life expectancy of 9.2 years of those patients OLDER THAN 50 YEARS is seen with early treatment with HAART.
- Whom do you treat? At what cost?

EVALUATION OF MEDICAL TESTS AND TREATMENT

COSTS

Decision tree

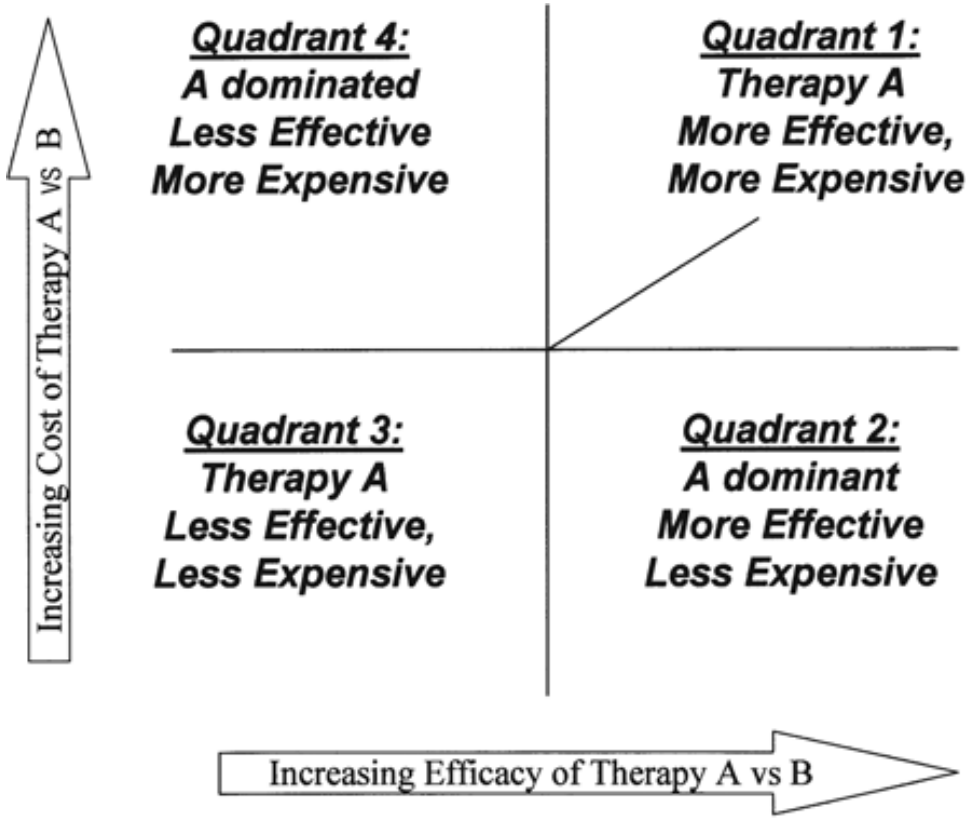


Source: Fuster V, O'Rourke RA, Walsh RA, Poole-Wilson P: *Hurst's The Heart*, 12th Edition: http://www.accessmedicine.com

Fig. 111-2 Accessed 08/10/2010

P, probability of health state; U, utility of health state.

Cost effectiveness



Source: Fuster V, O'Rourke RA, Walsh RA, Poole-Wilson P: *Hurst's The Heart*, 12th Edition: http://www.accessmedicine.com

incremental cost-effectiveness ratio in cost per life of quality-adjusted life-year gained.

Fig. 111-4 Accessed 08/10/2010

White counts as a screen

- Fewer than 0.5% of asymptomatic patients whose primary disease is not associated with leukocytosis will have an abnormal white blood cell count.
- Among patients whose total white blood cell count is normal, a white blood cell differential contributes to patient care in only 2.8%.

Hemoglobin as a screen

- The incidence of previously undetected hemoglobin abnormalities in women ranges from 6-13%.

- The incidence of previously undetected hemoglobin abnormalities in men rises to 6% after age 60.

Preoperative coagulation tests

- Screening preoperative coagulation tests are recommended by the American Society of Anesthesiologists for patients with Class I or II physical status only in the presence of hepatic or renal disease or with the inter-current use of anticoagulants.

Preoperative urinalysis

- Urinalysis is recommended as a screening tool preoperatively by the American Society of Anesthesiologists in patients in Class I and II physical status with hepatic or renal disease, diabetes, or in the elderly.

Preoperative chemistry tests

- The American Society of Anesthesiologists recommends preoperative chemistry laboratory tests for Class I and II physical status only if the proposed operative procedure is associated with known significant blood loss.

Preoperative tests of renal function

- Electrolytes and creatinine or BUN are the minimum recommended tests in patients with renal disease; or if taking diuretics or digoxin.

- Elderly patients are more likely to have renal impairment. Creatinine clearance is a better evaluation of renal status than serum creatinine or BUN.

Other preoperative tests

- With steroid use or in patients with diabetes, glucose and electrolytes are the minimum recommended tests.

- Liver enzymes, glucose, electrolytes, creatinine and BUN are recommended tests in those with liver disease.

Routine thyroid testing

- Patients with a family history of thyroid disease
- Asymptomatic patients older than 60
- Perimenopausal women
- Pregnant women older than 35 as well as those post-partum
- Diabetics

Routine thyroid testing

- Patients with autoimmune disease
- Patients with new onset of dementia or psychiatric disease
- Patients with new onset of heart disease
- Patients with obstructive sleep apnea

Pap smear as a screen

- Sexually active women older than 21
- First intercourse earlier than 18
- More than six sexual partners
- Oral contraceptive use for more than 10 years

- Screen throughout active sexual life
- Post-hysterectomy there is no cervix and screening is not necessary.

STD screen

- History of genital warts
- Partner with penile cancer or whose previous partner has had cervical cancer
- Female homosexual activity transmits HPV.

- Chlamydia screens are only for those less than 24 years of age or pregnant and at high risk. This is the group highly likely to be infected.

Screening for diabetes mellitus type II

- Early detection may not alter survival.
- Begin as early as 18yo if family history or obesity.
- Screen at any age if coronary disease, polycystic ovaries, or gestational diabetes present.
- Screen every 3 years with fasting glucose.

Screening for dyslipidemia

- Begin screening men older than 35 and women older than 45 years old if no heritable lipid disorder or known cardiovascular disease; else, begin screening at 25 years of age.
- Total cholesterol, HDL best tests
- Repeat every 5 years or with lifestyle change.

Screening for cardiovascular disease

- Begin blood pressure screening at 3 years of age.
- Abdominal Ultrasound in men who have used tobacco and are >65 years of age to detect abdominal aortic aneurysm. If negative, do not repeat.
- Thallium stress exercise testing after age 40 in women to evaluate chest pain (all others, after age 50).

Osteoporosis

- DXA only test with clinical correlation.
- Screen white, Asian women if >65 years old.
- Medicare permits repeat every 2 years.
- There are no data to suggest screening (and intervention to correct bone density) affects fracture rate.
- Fractures rare in men, women of Sub-Saharan origin.
- Screen earlier if postmenopausal AND with nutritional disorder OR steroid use for more than 60 days.

Screening for breast cancer

- Screen all women beginning at age 50.
- Repeat every three years if negative.
- May cease screening at age 75.
- Screening not recommended if life expectancy is less than 10 years.
- Begin screening earlier if first degree relative with breast or ovarian cancer.
- Consider BRCA gene testing.

Screening for prostate cancer

- <u>There are no data suggesting survival improved through early screening.</u>
- PSA at 50 years of age. If <3.0, repeat every 3 years; If 3.00-4.99, repeat every year. 25% of men with normal levels will have cancer; however, only 2% of these will be high grade.
- Begin screening at 40 years of age if of Sub-Saharan ancestry or if first degree relative with prostate cancer.
- 44% of men will be over- diagnosed with these parameters.

Screening for prostate cancer

- If initial PSA <1.00 AND >65yo, repeat screening not necessary.
- 85% cancers curable if found when PSA <5.0
- PSA velocity >0.5ng/yr is an indication for biopsy as it is associated with increased risk of cancer death over a follow-up period of 10-15 years.
- For those patients with negative biopsies but rising PSA, consider genetic testing. It is unlikely the patient will consent to a second round of biopsies.

Screening for colon cancer

- Screen all asymptomatic patients >50 years old.
- Screen earlier if first degree relative with colon cancer or polyps. Consider gene study.
- Immunochromatographic fecal occult blood test annually

  AND Flexible Sigmoidoscopy

  OR Double Contrast Barium Enema every 5 years.

  Six common fecal occult blood tests annually may be sufficient if the imunochromatographic method is not available.
- Colonoscopy only for positive screens; repeat every 10 years if negative

Screening costs to treat one patient

- Liver Transplant $234,000
- Mammogram before age 50 232,000
- Cancer of the prostate 146,000
- CABG, 2 vessel, for angina 106,000
- Captopril for hypertension 82,000
- Thiazide for hypertension 23,500
- Smoking cessation suggestion 1,300

Cost of prevention

H. Influenzae type b vaccination of toddlers	**Saves money and lives**
One time colonoscopy for those men 60-64 yo	**Saves money and lives**
Intense program of tobacco use prevention in 7^{th}-8^{th} grades	**$ 23,000/QALY**
Screening all over 65yo for diabetes mellitus, not just those with hypertension	**$590,000/QALY**

Number needed to treat or harm (cost)

- To calculate the cost of an intervention, one must have the number needed to treat, the length of time needed to treat in order to see a benefit, and the cost of the intervention.
- If 200 is the number needed to treat, 3 years is the length of time needed to treat in order to see a benefit, and the cost of the intervention is $1/day, then

  200 x 3 x 365 = $21,900

HIV treatment results

- Early treatment of HIV infection with HAART in patients OVER 40 YEARS OF AGE is associated with a life expectancy of 11.4 years if CD4 counts are >200cells/mm^3 AND viral loads are >300,000 copies/ml, rising to 12.9 years if CD4 counts are >500cells/mm^3 AND viral loads are <10,000 copies/ml.
- Little improvement in the life expectancy of 9.2 years of those patients OLDER THAN 50 YEARS is seen with early treatment with HAART.
- Whom do you treat?

Cost of intervention

Cognitive behavioral family intervention for patients with Alzheimer's	**Saves money**
Cochelar implants for all profoundly deaf children	**Saves money**
Implant cardioverter-defibrillator	**$ 52,000/QALY**
Immediate surgery in 70 year old man with newly diagnosed prostate cancer	**Adds to costs and loss of life**

Cost per quality adjusted life year

Strategy	Cost
Switch from tamoxifen to aromatase inhibitor in early stage breast cancer	**$22,900**
Impant cardioverter-defibrillator versus continued medical manazgement	**37,400-77,200**
Fusion surgery for degenerative spondylolisthesis with spinal stenosis versus conservative management	**120,000**
Trastuzumab for metastatic breast cancer versus standard chemotherapy	**150,000**
Erlotinib for pancreatic cancer versus gemcitabine alone	**370,000-500,000**
Helical CT screening for lung cancer in 60 year old heavy smokers versus no screening	**2,300,000**

Weinstein, MC, Skinner, JA, Comparative effectiveness and health care spending – implications for reform. N Engl J Med 2010; 362:460-465