UNM COLLEGE of PHARMACY

UNM RADIOPHARMACY 1973 2003 30 YEARS OF EXCELLENCE

Evaluating the Medical Literature
II. Application to Diagnostic Medicine

Continuing Education for Nuclear Pharmacists
And
Nuclear Medicine Professionals

By

Hazel H. Seaba, R.Ph., M.S.

The University of New Mexico Health Sciences Center, College of Pharmacy is accredited by the Accreditation Council for Pharmacy Education as a provider of continuing pharmacy education. Program No. 0039-0000-14-181-H04-P 5.0 Contact Hours or 0.5 CEUs. Initial release date: 10/15/2014

-- Intentionally left blank --

Evaluating the Medical Literature
II. Application to Diagnostic Medicine

By
Hazel H. Seaba, R.Ph., M.S

Instructions:

Upon purchase of this Lesson, you will have gained access to this lesson and the corresponding assessment via the following link < https://pharmacyce.health.unm.edu >

To receive a Statement of Credit you must:

1. Review the lesson content
2. Complete the assessment, submit answers online with 70% correct (you will have 2 chances to pass)
3. Complete the lesson evaluation

Once all requirements are met, a Statement of Credit will be available in your workspace. At any time you may "View the Certificate" and use the print command of your web browser to print the completion certificate for your records.

NOTE: Please be aware that we **cannot** provide you with the correct answers to questions you received wrong. This would violate the rules and regulations for accreditation by ACPE. The system will identify those items marked as incorrect.

Disclosure:

The Author(s) does not hold a vested interest in or affiliation with any corporate organization offering financial support or grant monies for this continuing education activity, or any affiliation with an organization whose philosophy could potentially bias the presentation.

This lesson is a reprint of that initially released in 1999. The content is judged still relevant and useful to the basic understanding of relevant literature documenting scientific clinical trials involving diagnostic agents.

EVALUATING THE MEDICAL LITERATURE
II. APPLICATION TO DIAGNOSTIC MEDICINE

STATEMENT OF LEARNING OBJECTIVES:

The purpose of this lesson is to provide nuclear pharmacists with educational materials appropriate for a thorough understanding of the technical and diagnostic performance of a diagnostic test. The educational goal is that, upon successful completion of this course, the reader will have obtained the knowledge and skill to use criteria to analyze a diagnostic test study, to synthesize criteria for the evaluation of specific diagnostic tests and to evaluate the quality of published studies investigating diagnostic test performance.

Upon successful completion of this lesson, the reader should be able to:

1. Given the sample values for a diagnostic test's results in a disease free group, calculate the reference or "normal" range of values.
2. Summarize and explain the relationship between normal, abnormal, diseased and desirable diagnostic test values.
3. Explain the rationale and limitations of using a 'gold standard' test to assess the utility of a new diagnostic test.
4. Given experimental data, calculate sensitivity, specificity, predictive values, and likelihood ratios of positive and negative test results.
5. Contrast the experimental and clinical utility of the sensitivity, specificity and predictive values of a diagnostic test.
6. Illustrate the information advantage of likelihood ratios.
7. Contrast "stable" properties of a diagnostic test to the "unstable" properties.
8. Relate the predictive values of a test to prevalence rates of the disease.
9. Given experimental data, construct a ROC curve and explain its purpose in establishing the cutpoint between reference and disease values.
10. Given patient and diagnostic test data, calculate the posterior probability for a disease.
11. Identify the methodological features necessary for an appropriate clinical evaluation of a diagnostic test.
12. Evaluate published research reporting on the diagnostic discrimination of a test or procedure.

COURSE OUTLINE

EVALUATING THE MEDICAL LITERATURE
II. APPLICATION TO DIAGNOSTIC MEDICINE

Hazel H. Seaba, R.Ph., M.S

INTRODUCTION

Contained within each patient is the information needed to determine his or her health status. Our ability to access this information - the patient's internal health database - describes the art and science of diagnosis. Appropriate clinical management of the patient rests on our ability to mine the patient's internal health database. Uncovering the information we need requires choosing the correct place to look for information, using the most appropriate tool and the ability to sift diagnostic pay dirt from slag. In this continuing education lesson we will consider the tool, that is, the diagnostic test procedure. Information theory provides methods to assess the quality of the information gained from the diagnostic test procedure. Decision theory provides the mechanism to translate the results of the diagnostic test into meaningful patient health knowledge.

This lesson builds on an earlier course, "Evaluating The Medical Literature I. Basic Principles" (Volume 17, Lesson 5). To fully benefit from this lesson, the reader may wish to review the earlier material. While much of the material in this lesson applies to only diagnostic test assessment research, material from the Basic Principles lesson ·applies to all clinical research study designs.

FUNDAMENTAL PRINCIPLES OF DIAGNOSTIC TESTING

Diagnostic test procedures add facts to the patient's health information repository that we are creating. At some point, decisions about health status, disease presence or absence and choice of treatment options will be made based on the patient's accumulated health information. More often than not, these decisions will be made with information that is incomplete or, worse yet, with information that is misleading or inaccurate. The function of the diagnostic test is to improve the quality of medical decision making and decrease the amount of uncertainty that surrounds each decision.[1] Diagnostic test results build upon the information gathered from the medical history and physical examination.

> "... diagnosis is not an end in itself; it is only a mental resting-place for prognostic considerations and therapeutic decisions, and important cost-benefit considerations pervade all phases of the diagnostic process."[1]

A patient's health outcomes, including economic, clinical and quality of life outcomes, are at least partially dependent upon the strengths of data in that patient's health information repository. However, linking diagnostic test procedures to the patient's health outcomes is tenuous. The framework for this association was first presented by Fineberg, et al[2] and developed further by Begg[3] and Mackenzie and Dixon[4] in the context of assessing the effects of diagnostic imaging technology on the outcome of disease. The economic impact of diagnostic procedures on society is also of considerable importance and has been added to the original model as the sixth level in the framework.[5] The six hierarchical levels in the framework are:

(1) technical performance of the test [reliability],
(2) diagnostic performance [accuracy],
(3) diagnostic impact [displaces alternative tests, improves diagnostic confidence],
(4) therapeutic impact [influence on treatment plans],
(5) impact on health [health-related quality of life], and
(6) societal efficacy.

Published evaluations of diagnostic procedures most frequently fall into level one or level two of the hierarchy. Assessments at level three through six are more difficult and fraught with design challenges. This lesson will focus on level one and level two assessments.

The "Perfect" Diagnostic Test

The real world in which we practice does not contain "perfect" tools. Before discussing the evaluation of diagnostic tests that we know will never achieve perfection, it is useful to consider the characteristics of an ideal diagnostic test. In their section on testing a test, Riegelman and Hirsch[6] describe the ideal diagnostic test as having the following attributes:

"(1) all individuals without the disease under study have one uniform value on the test,
(2) all individuals with the disease under study have a different but uniform value for the test,
(3) all test results coincide with the results of the diseased or those of the disease free group."

In the ideal world we would not only be able to perfectly discriminate between individuals with and without the disease, we would never encounter ambiguous test results. The test would be reliable,

irrespective of the testing environment or the operator, and provide accurate results regardless of the patient subgroups tested. Pragmatically, faced with less than ideal diagnostic tests, we can quantitatively estimate a test's ability to discriminate between diseased and disease free patients as well as estimate its reliability and accuracy under a variety of conditions.

ESTABLISHING THE TECHNICAL AND DIAGNOSTIC PERFORMANCE OF A TEST

Our evaluation of a diagnostic test involves establishing how close the test comes to meeting the expectation of identifying those individuals who do have a given disease and distinguishing them from those who do not have the disease of interest. The three variables of the evaluation are then: the disease free population, the diseased patient population and the test itself. Riegelman and Hirsch considered the evaluation of a diagnostic test to be, "largely concerned with describing the variability of these three factors and thereby quantitating the conclusions that can be reached despite or because of this variability."

The Disease Free Population: Normal or “Reference" Range

If all individuals who were free of the disease had the same blood level of endogenous chemicals and compounds, elicited the same response to external stimuli and looked exactly alike when viewed with medical imaging techniques, we could use these values to define the status of being free of the disease. Biological variability assures us that these potentially useful diagnostic values will not be the same in all disease free individuals, and in fact, are likely to be widely distributed over a continuum of values. Individuals free of the disease will generate a range of values for any given diagnostic test. This range of values for disease free individuals is called the *reference range*. In the past this range has been called the range of normal values. "Normal" misrepresents the range in that individuals with values within the range are not all healthy or free from disease, and secondly, the distribution of values may not be Gaussian (normal).[7]

Diagnostic test results represent the identification or measurement of some object or definable property. Measurements have four scales: nominal, ordinal, interval, and ratio. [8] Values are interval or ratio measurements when the numerical distance between the individual values is equal, each interval represents an equal amount of the quantity being measured, and there is a zero in the scale. Many diagnostic test results are interval or ratio scale. Less commonly, diagnostic test results are represented on the nominal scale. Nominal scale is named values, such as sex (male or female), hair color (brown, black) or race (white, Native American, Asian). Ordinal measurement scale represents a rank ordering

of values, for example, good, better, best. Numbers may be used to quantitate a property or ordinal scale, such as a ten point scale for pain intensity. However, statistical manipulation of ordinal scale numbers may be limited as the interval between the numbers may not be equal and does not necessarily represent an equal quantity of what was measured, in this example, pain. Diagnostic test result values are also classified as being either continuous or dichotomous. Continuous values have the properties of being at least ordinal or higher scale and fall within some continuous range of values, for example, values of left ventricular ejection fraction from a gated blood pool procedure. Dichotomous values are categorical, a kind of nominal measurement representing the presence or absence of something. Diagnostic test results are dichotomous when the patient either has this property or does not have the property, such as visualization versus non-visualization of the gallbladder in a hepatobiliary imaging procedure. Continuous scale test results may be reduced to a dichotomous scale, such as disease present or disease absent.

The reference interval is constructed by measuring the diagnostic test values in individuals who are believed to be free of the disease. The reference sample tested is generally a convenient group of individuals (such as students, healthy volunteers, clinic employees, hospital staff) who are assumed to be free of the disease. Other diagnostic tests and examinations may be done on these individuals to establish their disease free status. Ideally the reference sample would represent a wide range of disease free individuals of both sexes, from all age groups, and with ethnic diversity.

The reference range of values for a diagnostic test is most frequently defined as the central 95% of the values of healthy individuals. If a range of values also exists for individuals known to have the disease, other methods of declaring the reference range may be used, including: use of a preset percentile, use of the range of values that carries no additional risk of morbidity or mortality, use of a culturally desirable range of values, use of a range of values beyond which disease is considered to be present, use of a range of values beyond which therapy does more good than harm.[9]

A reality of using the central 95% of values of healthy individuals is that 2.5% of individuals at each end of the range, known to be healthy, will be identified as outside the reference range. In clinical practice it is important to remember that just because a diagnostic test value falls outside of the reference range, it does not necessarily mean that the individual has a disease.

A frequency analysis of the test results from the reference sample will establish whether the results have (or can be transformed into) a Gaussian distribution or not. For a Gaussian distribution, the central 95% can be calculated at the mean, plus or minus two standard deviations. If the results are not Gaussian, a nonparametic analysis can sort the values from lowest to highest value and exclude the lowest 2.5% and highest 2.5% of the values.

If there are significant differences among a population that effect the diagnostic test results, the reference sample may be restricted to just one group. Age, sex, race, and smoking status frequently represent legitimate subsets' of test values.

Variability of the Diagnostic Test

When we consider the results of a diagnostic test procedure in a specific patient, we would like to be sure that the test is measuring what we think it is measuring and that any deviation in the test's results from the reference range values or from prior values of this test in this patient, is due only to the disease process or a change in the disease process in this patient. This is analogous to the evaluation of a new drug in a clinical trial - we would like to be confident that the outcome we measure in the study subjects is due to the new drug and not due to any other variable, such as the disease wanes on its own, the subject's general health improves, the individual measuring the drug's response in the patient is inconsistent, the instrument recording· the outcome is failing or any of a multitude of other random and non-random events (biases) that plague clinical research.

In the context of experimental designs, Campbell and Stanley [10] identified twelve sources of variability that threaten the validity of an experimental design. The twelve factors are relevant to our current discussion on two counts, first, one of the twelve factors is *'instrumentation* ' and secondly, when we compare one diagnostic test to another within a clinical experiment, all twelve of these factors need to be controlled to establish the validity of the comparison. Instrumentation has two facets: the test (or testing instrument) itself and the operator. Any change in the test itself, for example, calibration, may result in incorrect test results. The person or persons who calibrate the test, apply or administer the test, observe the results and record the results have the opportunity at each of these steps

Factors Jeopardizing Internal and External Validity:[10]

1. History
2. Maturation
3. Testing
4. 'Instrumentation
5. Statistical regression
6. Selection bias
7. Experimental mortality
8. Selection-maturation interaction
9. Interactive effects of testing
10. Interactive effect of testing
11. Interaction of selection biases and experimental variable
12. Multiple-treatment interference

to perform the step incorrectly or inconsistently and thus invalidate the test results. Analogous to instrumentation, radiopharmaceutical quality (especially radiochemical purity) must be controlled in order to prevent incorrect test results from variation in biodistribution. Of the twelve sources of invalidity, instrumentation is one of the easier factors to control. Operator training and good laboratory techniques can decrease instrumentation bias. Under experimental conditions, if all other sources of variation are controlled, the intrinsic stability, dependability and value of the test itself will be measurable. The variability of the test itself should be small compared to the variability of the range of normal values for the test, otherwise, the test will mask differences in values due to biological factors.

Reproducibility of the Test

Reproducibility is synonymous with reliability. A test's reliability can be judged by determining the ability of a test to produce consistent results when repeated under the same conditions. Re-testing a diagnostic test requires the laboratory, patient and operator/observer to remain the same during each test session. Both intra- and inter-observer variability are possible and should be considered. Re-testing must also be done in such a manner that the results of the first test are not known during the re-test.

The precision of the measurements is a reflection of reliability of the data/measurements themselves. Precision is the agreement of the measurements with one another and is frequently described by the range of the results or their standard deviation.

When re-testing is done, it is always possible that some of the agreement between measurements or individuals is due to chance. Kappa (ƙ) is an index of agreement that corrects for chance agreement. Kappa incorporates both the observed proportion of agreement between two measures/observers and the proportion of agreement expected due to chance.[11] If the agreement between two observers is perfect, Kappa's value is +1.0, disagreement between observers can reach -1.0. If the agreement between observers is no different than that expected by chance, the value is 0. A Kappa value of ≥0.75 implies excellent reproducibility.

In addition to test/retest reliability, split-half reliability may also be assessed. Split-half reliability evaluates the internal consistency of one part of the test with other parts of the test. For example, the questions in a survey instrument may be compared for redundancy or congruency using split-half reliability.[12]

Based on an investigation that corrected for chance agreement in data, Koran identified several factors that contribute to increased reliability of measurements.[13] These factors include: having a high proportion of 'normals' in the evaluation, high observer training, consideration of only a small number of diagnostic categories, abnormal values that are severe, observations that are dichotomous rather than continuous in scale.

Accuracy of the Test

Accuracy describes the correctness of the test values, that is, the agreement of the test value with an independent judgment that is accepted as the true anatomical, physiological or biochemical value. Accuracy requires reliability. However, the converse is not true, measurements can be reliable without being accurate. In addition to being reliable, measurements also need to be freed of systematic tendencies to differ from the true value in a particular direction. Systematic error is bias. In our discussion of establishing the diagnostic performance of a test, we will consider several sources of bias.

Under experimental conditions, the accuracy of a test can be established by comparison with an artificial sample ('known') or a phantom image. This is sometimes referred to as experimental accuracy. In clinical practice, of course it is very difficult to know the patient's absolute diagnostic truth. The quandary of assessing the accuracy of any diagnostic test (clinical accuracy) is having a true value for comparison.

Validity of the Test

The validity of a diagnostic test is distinct from the reproducibility and the accuracy of a test. Validity asks the questions, "Are we measuring what we think we are measuring?"[8] A valid test is one that is appropriate for the diagnostic question being asked.

Three types of validity are most frequently considered: content validity, criterion-related validity and construct validity. Content validity is a judgment of whether or not the test is representative of what is supposed to be measured. Criterion-related validity is established by determining whether or not the test results agree with the results of one or more other tests that are thought to measure the same anatomical, physiological or biochemical phenomena. Construct validity seeks to find out why, theoretically, the test performs as it does. What is the real biological property being measured that explains why the results of the test very among individuals?

Variability of the Diseased Population

Although we most frequently refer to individuals as either having or not having a given disease – a dichotomous classification of decease – for almost all diseases, the disease process is continuous. Over time the disease severity and the number of signs and symptoms of disease escalate. In general, it is more difficult to diagnose a disease in its early stages than in its later stages, when the disease process is stronger and more distinct. Not only are there likely to be continuous changes in the disease manifestations over time that will complicate diagnosis, but there are also likely to be other potentially confounding variables present in the patients. Patient variables that may make a difference in the disease presentation include sex, age, presence of other diseases, nutrition status and current drug therapies. In brief, we can expect a wide variability of response to a specific diagnostic test from individuals who do have the disease.

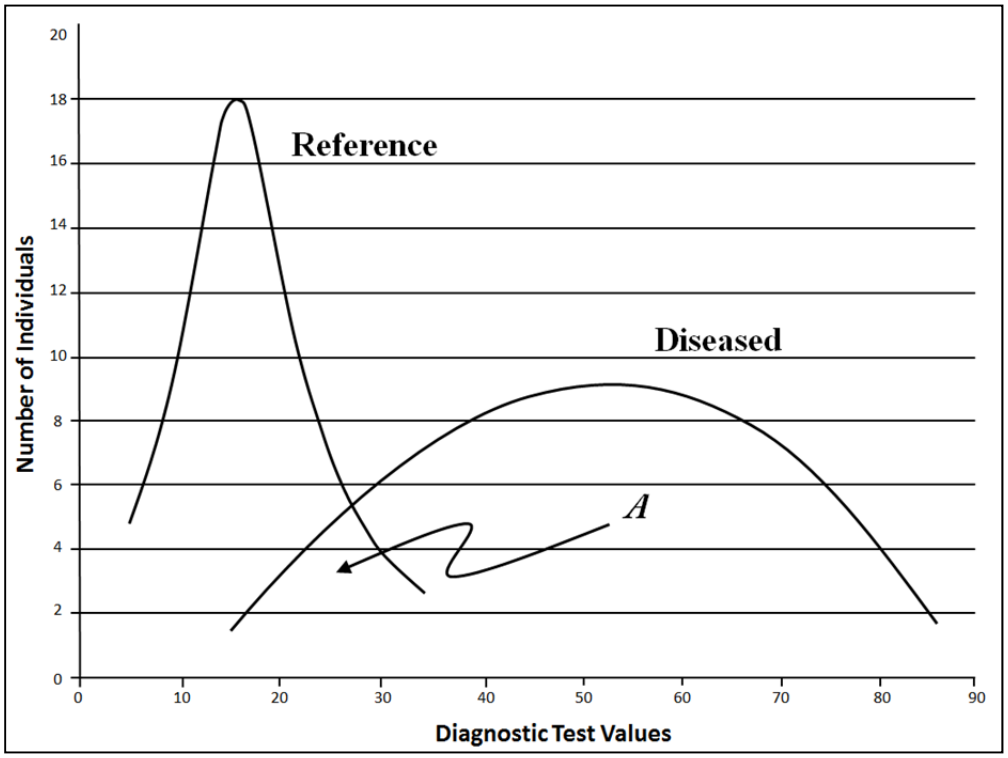

Figure 1. Diagnostic Test Results

Ideally, the variability of the diagnostic test in the disease free reference population and that of individuals with the disease will not overlap. In practice however, the reference population will frequently have results values in common with diseased individuals. In Figure 1, this is the area indicated by ***A***. Which test value should be used as the cut point to delineate individuals who are free

of the disease from those with the disease? Even though we are aware of this area of equivocal values between the reference group and the disease group, it is important to establish a cut point (discrimination limit). Earlier we mentioned that if the range of values for individuals with the disease was known, the cut point between the reference group and the diseased group could be established using criteria other than assigning the central 95% of reference values to the disease free group.[9] We will consider these other methods in the medical decision making section of this lesson.

Defining Disease: The Gold Standard

By definition, the gold standard diagnostic test is that test whose results determine the subject's disease status. Any test assigned gold standard status is accepted to be 100% accurate. Gold standard status is a clinical judgment. Historically, autopsy[12] and biopsy have been used as gold standards for diagnosis. While an autopsy may provide an unequivocal diagnosis for a research study, pragmatically, it cannot be used in clinical practice. For a given disease condition, generally, the current best diagnostic test becomes the gold standard. While the current gold standard test for a disease may represent the best, most practical test we have, it also may be "gold" in name only. For many disease conditions, truly accurate diagnostic tests do not exist and the choice of which test to assign gold standard status to is not clear. Difference of opinion will exist.

It is against the gold standard that a new diagnostic test will be compared to determine its accuracy. However inadequate a gold standard might be, it is necessary to assess a new diagnostic test procedure against the best current test. It is not sufficient to determine whether or not the new test is more frequently associated with discovering disease than chance alone. This is analogous to the evaluation of a new drug. Regardless of whether the new drug is compared to a placebo control or to the current drug of choice (active control), it is the objective, unbiased comparison that estimates the drug's effectiveness either benchmarked to the placebo or to standard therapy.

Since the gold standard is considered 100% accurate, the new diagnostic test can not outperform the standard. The accuracy of the new test swill always be less than or equal to the gold standard. In truth, we frequently expect that the new diagnostic test will be an advancement over the current hold standard test. With clinical use, the new test may convince practitioners of its superiority, allowing it to eventually become the gold standard.

PROCEDURAL STEPS TO DETERMINE DIAGNOSTIC PERFORMANCE

Diagnostic performance described the ability of the test to correctly determine which individual have the disease condition and which do not. Clinical studies with the goal of establishing the diagnostic performance of a test share some of same design features as therapeutic efficacy controlled clinical trials. The diagnostic performance design should ensure a study setting that provides a fair, unbiased assessment of the test's accuracy using the gold standard benchmark.

Technical Performance

High technical performance (reliability) of the test provides a favorable foundation for high diagnostic performance by the test. Reliability or reproducibility of the test under a variety of clinical conditions by different observers or operators is essential. If the images produced by a new test are inconsistently interpreted by the same radiologist or do not receive the same interpretation by different radiologists, the test is not useful. To ensure reliability, the protocol for the evaluation of a new diagnostic test should include training with standard operating procedures for test operators and interpreters. Standardized procedures can minimize irregularities in radiopharmaceutical quality, sample collection, instrumentation, data collection and recording. The agreement both within and between the individuals who execute the test and those who read the test should be assessed. The variability of the instrument under controlled conditions should also be assessed. The goal of the study design is to eliminate any source of variability from test itself (instrument), the operator or the interpreter, so that the experimental variability of the difference between the new test and the gold standard is not masked and can be measured.

Riegelman and Hirsh[6] identified five basic steps to determine diagnostic performance:

1. choose the gold standard test,
2. perform the gold standard test on a full spectrum subjects,
3. test all subjects with the new diagnostic test,
4. record and compare the results of the new test with the gold standard test in a two by two outcome table, and
5. calculate the proportions of accurate and inaccurate results for the new test.

One of the seven criteria used by Reid, et al[15] to determine the quality of diagnostic test evaluations is whether or not the reproducibility of the test is shown by some measure of observer variability and/or instrument variability. The found that only 23 percent of the 112 studies reviewed provided evidence of reproducibility with either percent agreement of observers or kappa statistics and only 25 percent of studies reported interassay and/or intra-assay coefficients of variation. Bringing this criterion into an individual clinical practice, Jaeschke, et al[16] ask the questions, "Will the reproducibility of the test result and its interpretation be satisfactory in my setting?"

Test Reproducibility Criterion:[15]

- If observer interpretation is used, some of the test subjects are evaluated for a summary measure of observer variability,
- If no observer interpretation is used, a summary measure of instrument variability is provided.

Gold Standard Choice

The choice of which diagnostic test to use as the gold standard is a clinical decision. Realistically, the gold standard choice may be a compromise between a testing procedure that provides the most accurate result and a procedure that is less invasive or less costly. In the introduction to an article, investigators generally present the justification for their choice of reference standard. Ideally the gold standard not only represents a test that makes sense clinically – that is, if the new test is successful, it would replace or provide an alternative to the gold standard for this particular disease – but also, the test does accurately classify patients as either disease free or diseased. As the new test's results are compared to the gold standard's accuracy, errors in classifying patients by the gold standard will perpetuate themselves in the assessment of the new test's accuracy.

What makes a 'good' gold standard test? According to Arroll, et al. [17] a well-defined gold standard is either:

- Definitive histopathologic diagnoses (autopsy, biopsy, surgery),
- Standard diagnostic classification system (for example, American Psychiatric Association *Statistical and Diagnostic Manual for Mental Disorders* for depression), or
- Results of other well-established diagnostic tests, if explicit criteria are given for when the target disease is said to be present.

In Arroll, et al's review of 126 selected clinical journals published in 1985, 88% of the diagnostic test articles used a well-defined gold standard.

In the FDA's 1999 final rule on "Regulations for In Vivo Radiopharmaceuticals Used for Diagnosis and Monitoring," the phrase 'gold standard' is not used; however, an expression with the same meaning is: "The accuracy and usefulness of the diagnostic information is determined by comparison with a reliable assessment of actual clinical status. A reliable assessment of actual clinical status may be provided by a diagnostic standard or standards of demonstrated accuracy. In the absence of such diagnostic standard(s), the actual clinical status must be established in another manner, e.g., patient followup."[18] The FDA considers a comparator or clinical follow-up necessary to establish the accuracy and usefulness of a radiopharmaceutical's claim for detection or assessment of a disease or a pathology. As an alternative to a gold standard diagnostic test results, clinical follow-up for an adequate period of time can be used to establish the patient's true disease status. Unfortunately for many diseases the time needed for the disease to progress to the point where the diagnosis is unequivocal by direct observation may be quite long.

The gold standard test, by definition, is assumed to be perfect with no false positive or false negative test results. If this assumption is not true, the false negative rate (1 – sensitivity) and the false positive rate (1 – specificity) of the test being evaluated are overestimated. [Author's Note: Sensitivity and specificity are defined and discussed in detail in section 'Medical Decision Making Applied to Diagnostic Test Performance,' page 23, and in Table 1, page 41.] This possibility was investigated by Line, et al.[19] These investigators calculated the sensitivity and specificity of antifebrin scintigraphy, ^{99m}Tc-antifibrin (^{99m}Tc-T2G1s Fab'), in patients with suspected acute deep venous thrombosis (DVT) using two different methods. Two patient groups were studied, one with low DVT prevalence and one with high DVT prevalence. The first method compared antifebrin scintigraphy to contrast venography, the gold standard assumed to have no error. The second method calculated sensitivity and spec8ificity using a maximum likelihood procedure that does not include comparison to a gold standard. The maximum likelihood procedure uses diagnostic test results from two populations with different disease prevalence. Sensitivity and specificity of antifebrin scintigraphy as estimated by comparison to the venography gold standard were substantially lower than those calculated with the maximum likelihood procedure. The authors concluded that a gold standard with errors will bias the sensitivity and specificity of the test being evaluated downward and that this effect was operating in their study. They suggested that contrast venography may not be a good gold standard for DVT.

There are techniques available to decrease or minimize the bias introduced into a diagnostic test evaluation by an imperfect gold standard.[20] One technique involves making it less likely to diagnose a

patient as disease present when the patient is truly disease positive and making it less likely to diagnose a patient as disease free when the patient is truly disease negative. Thus the test is more likely to find only true positive individuals as positive and true negative individuals as negative. This can be accomplished by using a rigorous definition of disease when the test's sensitivity (ability to diagnose a positive patient as diseased) is evaluated and likewise using a lax definition of the disease when the test's specificity (ability to diagnose a negative patient as disease free) is measured. Another technique recognizes that when the gold standard misclassifies patients, both sensitivity and specificity are influenced by the disease prevalence. Prevalence is the proportion of individuals who have the disease at a given time. If the disease prevalence is high, it is easier to diagnose positive individuals; if the disease prevalence is low, it is easier to diagnose disease negative individuals correctly. If sensitivity and specificity are assessed in both high and low prevalence environments, bias may be minimized. It may also be helpful to use a surrogate for a positive and negative disease diagnosis. For example, instead of diagnosing the presence or absence of the disease in an individual, it may be possible to use the diagnostic test to determine if the patient will respond favorably to drug therapy for that disease or whether the patient will be a drug therapy failure. Lastly, mathematical corrections can be applied to test results known to have error.

Selection of Study Sample

At the point a new drug is approved for marketing by the FDA, the clinical experience with the drug may be limited to only a few thousand or even a few hundred study subjects. Frequently what we know about the drug's safety and efficacy not only changes and increases, but also improves in accuracy as more knowledge of the drug is gained through its use in a broader patient population under a variety of conditions. Similarly, an evaluation of a new diagnostic test is most credible if the study or studies of the test's accuracy includes a broad selection of individuals both with and without the targeted disease.

If investigators were able to study the new test in a group of subjects suitably large enough that the group contained all the conditions under which the disease occurs and does not occur, the measurement of the new test's accuracy would dependably reflect the entire disease state (provided the gold standard was accurate). T])e investigator rarely, if ever, has the ability to study a census of individuals with a disease. Almost equally unlikely is the opportunity to randomly sample the population made up of disease negative and disease positive individuals. Even if random sampling of the population was possible, the number of disease positive individuals in the population would be very small compared to the number of disease negative individuals. The sample drawn would most likely contain few disease

positive individuals. Thus the sensitivity value (fraction of disease positive individuals successfully identified) would have a wide confidence interval.[21] To improve the estimate of sensitivity, a non-random, disproportionate sample is used. It is still important that the disproportionate sample represent as many characteristics of individuals with and without the disease as possible. The range of characteristics represented in the sample is called the spectrum. Under circumstances of wide subject spectrum, the new test will receive a fair challenge of its abilities to correctly diagnose a variety of subjects.

Spectrum includes pathologic, clinical and comorbid components.[22] Most disease conditions have a wide spectrum of pathologic features, such as extent of the disease, location of disease, cell types. If only patients with a given tumor size (pathology) are used in the diagnostic test evaluation, the usefulness of the test in patients with smaller or larger tumors will not be known. The clinical features of the disease describe the chronicity and severity of symptoms. Either of these features can influence the test results. Comorbidities of the patient may also affect the results of the diagnostic test. In patients who are truly disease negative, the goal for the diagnostic test is to minimize false positive results. A challenging spectrum of disease negative subjects would include subjects with different pathologies in the same anatomical area as the target disease, individuals with physiologic disorders that affect the same organ as the target disease or prevent the proper distribution of a test substance or imaging agent, and subjects with other diseases that may interfere with diagnosis. Unfortunately, to achieve better control in the study, the tendency is to choose disease free study subjects that have no comorbidities with disease symptoms that overlap those of the target disease.

The challenge in patients who are diseased is to diagnose as positive as many as possible and, at the same time, avoid misdiagnosing disease free individuals (a false positive diagnoses). To ensure that the new test will function properly in all individuals with the target disease, patients with a wide spectrum of the disease's pathologies should be included. Individuals who have been ill for long and short periods of time and who have mild and severe symptoms should be chosen. Also individuals who not only have the target disease, but other diseases as well should be represented in the study.

The sample spectrum should also include both genders and a variety of subject ages and ethnicity. It is possible that a given diagnostic test will perform the same in all subjects with and all subjects without the target disease. However, the only way to know if this is true is to study the diagnostic test in all manner of patients. If the diagnostic test is found to perform differently on different groups of

subjects, spectrum bias is said to exist. In addition to choosing a broad spectrum of study subjects, the investigators, in order to assess spectrum bias, must also analyze the study's results by patient subgroups. Subgroup analysis can pinpoint age groups, phases of disease, or other subsets of patients who will have accuracy estimates, that is sensitivity and specificity, different than the majority of patients with the target disorder. Spectrum bias is illustrated in Morise et al.'s[23] review of discriminant accuracy of thallium scintigraphy in patients with possible coronary artery disease. A derivation group received single photon emission computed tomographic (SPECT) and a validation group received SPECT and planar thallium-201 scintigraphy. They found differences in (1) sensitivity and specificity for the two separate study samples based on all defects verses reversible defects, and (2) sensitivity, but not specificity, based on the number of diseased vessels involved. The accuracy of exercise ECG was also found to be lower in women than men.

Analysis of Pertinent Subgroups Criterion:[15]
- Results for indexes of accuracy are cited for any pertinent demographic or clinical subgroups of the investigated population (e.g., symptomatic versus asymptomatic patients).

A broad spectrum of study subjects increases the confidence with which we can extrapolate the results of the diagnostic test to patients in our local practice. A narrow spectrum of study patients does not necessarily decrease the internal validity of the study - it may only affect the external validity of the study, that is, the study's generalizability. If our local patients are so similar to those in the study that they could meet the inclusion and exclusion criteria of the study, we are most comfortable using the test locally. But, regardless of however broad the study subject spectrum was, we frequently are considering local patients who are sicker than those in the study, are older or younger, or have a coexisting disease or diseases not included in the study. Under these circumstances it may be useful to ask questions of not only how different your patient is from those in the study, but also how similar your patient is to those in the study. If the diagnostic test uses a pharmaceutical agent, knowledge of the agent's pharmacology is very helpful to anticipate patient characteristics or pathologies that may interfere with the mechanism of action of the testing agent.

Spectrum Composition Criterion:[15]
- age distribution,
- sex distribution,
- summary of presenting clinical symptoms and/or disease stage, and
- eligibility criteria for study subjects are given.

Studies executed with a narrow spectrum of subjects are most likely to result in falsely high test accuracy.[22] The test's sensitivity and specificity will be overestimated. In their study of the application

of methodological standards in diagnostic test evaluations, Reid et al[15] found that only 27% of the studies they reviewed met their criteria for specifying the spectrum of patients evaluated.

Sample Size

Researchers, clinically evaluating diagnostic tests, construct a study design where the accuracy (sensitivity and specificity) of one or more diagnostic tests is calculated in subjects whose positive or negative disease status has been established with a gold standard diagnostic test. For convenience and precision considerations, the study sample is frequently selected disproportionately, that is, the proportion of disease positive individuals included in the sample is not the same as that in the general population. In many studies the subjects are selected from a group of patients whose disease status is known (patients have already undergone the gold standard diagnostic test). Under these circumstances the proportion of disease positive to disease negative subjects is under the investigators immediate control. This is usually not true in clinical trials of screen diagnostic tests. In a screening study of a large population, subjects with unknown disease status are consecutively enrolled into the study.

Precision of Results for Test Accuracy Criterion:[15]

- SE or CI, regardless of magnitude, is reported for test sensitivity and specificity or likelihood ratios.

The proportion of disease positive to disease negative individuals included in the study may be arbitrarily determined by the investigator, may be determined from a calculation of sample size based on a desired confidence interval [21, 25, 26], or may be determined by the number of subjects selected in a screening investigation. Ideally the sample size for the study is calculated prior to subject enrollment. The calculation will determine the number and optimum ratio of disease positive to disease negative subjects. Regardless of the method used to determine the sample size, it is important to remember that the proportion of disease positive to disease negative individuals in the study may be artificial and not reflect the disease's real prevalence in the general population.

The evaluation of a diagnostic test aims to measure the accuracy of the test. The outcomes measure are sensitivity and specificity of the test derived from comparison to the subjects' real (gold standard) disease presence or absence status. Sensitivity and specificity are point estimates. We would like the sensitivity and specificity values to be precise. The standard error (SE), or width of the confidence interval, measures the precision of these values. A 95% confidence interval is defined as an interval that contains the point estimate about 95 times in 100 replications of the study. [25] Thus in about 5 replications the point estimate would not be included in a 95% confidence interval. For large sample

sizes (n>0) the two-tailed 95% confidence interval for sensitivity or specificity can be estimated with the following equation: [15,27]

$$95\%\ CI = \text{point estimate} \pm (1.96) \times (SE)$$

The upper and lower values for the confidence interval provide us with an indication of how low and how high the real value could be and still be compatible with the data. Sample size influences precision. The larger the sample size, the more precise are the calculated sensitivity and specificity estimates, thus the narrower are the confidence intervals. If the investigator wishes to limit his or her error to a predefined level, a sample size that will generate a narrow confidence interval can be calculated. [24] We are also interested in how accurately the test identifies disease positive and disease negative individuals (the sensitivity and specificity of the test). Our ability to correctly measure the difference between the new test and the gold standard values is also dependent upon sample size. The larger the sample size, the less likely are we to incorrectly measure the two proportions (sensitivity and specificity). We know that the measurements generated by the study data are estimates of the true values and that we always have a chance of arriving at an incorrect value or conclusion. However, we would like the study to have enough power (80 to 90%) to determine, with a given degree of confidence, the size of the two proportions. Alternatively, the investigator may be interested in testing a hypothesis that the new diagnostic test has sensitivity and. specificity values that are no more than some specified distance from the gold standard test. The study's power, 1- beta (beta is Type II error), is related to the value of alpha (Type I error), variability of the events (i.e., disease presence), delta (the clinically meaningful values for sensitivity and specificity) and sample size[27]. We generally set the probability of making a Type I error (alpha) at 0.05; this corresponds to a 95% confidence interval (CI). A Type I error results when the investigator concludes that the two proportions (for example, sensitivity of the new test and that of the reference test) are statistically significantly different when they are not different. Beta values of 0.10 or 0.20 are desirable. A Type II error occurs when the investigator concludes that two proportions are not statistically significantly different when they are different. Clinically meaningful values of sensitivity and specificity are determined by the investigators - frequently these values are based on the investigator's best estimate of the test's sensitivity and specificity. The variability of events (disease prevalence in the target population) is chosen by the investigators, generally from the published literature. A sample size is calculated for a given sensitivity value and another calculated for a given value of specificity. The final sample size is the larger of these

two values. By choosing the larger value, the sample size will be adequate to estimate both sensitivity and specificity with the desired precision.

Linnet[26] has pointed out that for continuous scale diagnostic test values, it is important to also consider the sampling variation of the cutpoint (discrimination limits) between disease free and diseased. Sensitivity and specificity depend upon where the cutpoint is drawn. If the sampling variation of the cutpoint is not considered, the probability of making a Type I error may increase from 0.05 to values of 0.l0-0.35.

How do we know if the study has an adequate sample size? This is a difficult literature evaluation question. If the investigators provide the a priori calculated sample size, we at least know that they considered the variables necessary to calculate sample size. If one has mathematical skills, the power of the study can be calculated retrospectively using the sensitivity and specificity values calculated in the study and the best estimate of disease prevalence. Tables of sample sizes for given sensitivity and specificity values at 95% confidence intervals and 90% power are provided by Buderer.[27] One could look up the sensitivity or specificity values reported in an article in the Buderer tables, note the required sample size for the reported sensitivity or specificity, and then compare the needed sample size to that actually used in the article. Freedman[28] also provides a table to determine needed sample sizes for various sensitivity and specificity values. Otherwise, we are left with judging the adequacy of the sample size by considering the reasonableness of the number of subjects in the study and the study's results. Kent and Larson[29] recommend a sample size of 35 to several hundred patients for high quality studies of diagnostic accuracy.

Prevalence

Prevalence is a probability - it represents the number of people in the population who have a disease at a given point in time.[6, 30] The 'point in time' can be a specific point, such as June 6, 2000 (point prevalence) or a period of time, such as 2000 (period prevalence). In contrast, incidence is a rate representing the number of new disease cases that develop in the population over a unit of time. The two are related:

prevalence = incidence rate x duration of the disease

Prevalence tells us how many patients with a given disease are available to be diagnosed with the disease. Diseases with either a high incidence and/or a long duration will have a high prevalence.

Generally we believe it is easier to identify individuals positive for a disease if the disease's prevalence is high as the proportion of positive individuals in the population is large[20, 31] Thereby, our chance of encountering a positive patient is high. Diseases with low prevalence seem more difficult to diagnose, as the number of positive individuals in the population is small.

How is disease prevalence related to accuracy of diagnosis, that is, sensitivity and specificity? Sensitivity and specificity are considered by some authors to be independent of disease prevalence.[6, 20, 21, 32, 33, 34] Stable sensitivity and specificity values under different disease prevalences would be an advantage. Worldwide, disease prevalence varies from country to country. A test whose accuracy is the same regardless of the local disease prevalence would allow us to extrapolate sensitivity and specificity values from the primary literature to any practice site at any location. Under conditions of stable sensitivity and specificity, the accuracy of one test can be compared with the accuracy of competing diagnostic tests for the same disease.

Unfortunately, the relationship between sensitivity and specificity and prevalence is probably not complete independence. Literature as far back as the 1960's has pointed out examples of diagnostic test sensitivity and specificity for a disease varying by the population under study.[35] Sensitivity and specificity can vary with patient demographic characteristics, such as age or sex, and also with the clinical features of disease, such as severity, duration and comorbidities.[3, 15] In these examples, a possible explanation is that the prevalence of the disease is truly different for different ages and sexes or for different stages of the disease, severity of disease or in the presence of other pathologies. If the diagnostic test was evaluated in a wide spectrum of patients, sensitivity and specificity represents an "average" accuracy across all these variables. If a very narrow spectrum of subjects is tested or if the investigator analyzed the indexes by subgroups, sensitivity and specificity indexes apply to only those subgroups.

Coughlin and Pickle[31] point out that part of the agreement between the diagnostic test and the gold standard may be due to chance and offer a mathematical correction for chance agreement. Diseases with high prevalence are likely to offer more individuals with positive disease status in the sample and thus chance agreement may be a large factor in the sensitivity value. Likewise, diseases with low prevalence provide more individuals with negative disease status and thus chance agreement may be a large factor in the specificity value.

In her calculation of sample sizes adequate to estimate sensitivity and specificity with a given precision, Buderer [27] incorporates the prevalence of the disease in the target population. The number of subjects who are disease positive and negative is dependent upon the disease prevalence. Within the study, the total positive subjects (true positives and false negatives) and total negative subjects (true negatives and false positives) are the denominators for the standard errors (SE) of sensitivity and specificity. The width of the confidence interval (CI) is dependent upon SE. The width of CI influences the sample size. Thus the sensitivity and specificity indexes are influenced by disease prevalence.

Brenner and Gefeller[36] build on Buderer' s justification by pointing out that prevalence in the population is determined for most diseases by the diagnostic cutpoint. Regardless of how the cutpoint is established, there will be some disease positive individuals labeled as disease negative and conversely some disease negative individuals labeled as disease positive. This same cutpoint determines both, (1) the disease prevalence of the population, that is the number of diseased individuals in the population at a given time, and (2) the magnitude of the diagnostic misclassification of individuals at the time sensitivity and specificity of the test are determined. Thus disease prevalence and diagnostic misclassification are related. For example, the cutpoint could be adjusted to maximize sensitivity at the expense of specificity and this would also change the disease's prevalence. Most diseases are diagnosed by measuring some continuous variable patient characteristic and the above reasoning is applicable. However, if the diagnostic test truly measures a dichotomous variable, such as alive or dead, there is no cutpoint and sensitivity and specificity indexes are thought to be independent of prevalence.

Our concern for the influence of prevalence on sensitivity and specificity can be somewhat mollified if the spectrum of individuals in the diagnostic test study is clearly identified and, if the spectrum is broad, subgroup analyses on important disease groups are done. Other diagnostic indexes, such as predictive values, and other decision methods, such as receiver operator characteristic curves, are also available and provide a different view of diagnostic test discrimination.

Measurement and Data Collection

Despite the disadvantages of sensitivity and specificity, they are the primary measurements of diagnostic test efficacy that appear in the medical literature. Y erushalmy[37] first used the terms sensitivity and specificity to quantitate observer variability among radiologists. The terms actually have two meanings: analytical sensitivity and specificity and diagnostic sensitivity and specificity.[38] A

single laboratory test may have both analytical and diagnostic sensitivity and specificity. A laboratory test that is an assay (measures a substance) has an analytical sensitivity that is defined as the assay's ability to measure low concentrations of the substance or detect a change in concentration. If the target substance is also a surrogate for a disease condition, the assay may be used to detect the substance in a population to determine disease presence or absence. At this point, the test becomes a diagnostic test and the test's ability to detect individuals who have the disease (sensitivity) becomes relevant. While the diagnostic test has to be able to measure the target substance at meaningful levels or concentrations, it is also important that the diagnostic testing process obtains a patient sample that contains the target substance. A diagnostic test with high analytical sensitivity may have low diagnostic sensitivity if the target substance sampling procedure is inadequate.

Analytical specificity is defined as the ability of the test to exclusively identify a target substance, such as just the β (beta) subunit of human chorionic gonadotropin (HCG) rather than both the α (alpha) and β (beta) subunits of HCG. HCG (containing both α and β subunits) is produced by the placenta and its presence in urine or serum is used to diagnose pregnancy. Other hormones, such as luteinizing hormone, thyroid-stimulating hormone and follicle-stimulating hormone, also contain an identical ex subunit. Newer diagnostic tests specific for the β subunit decrease false positive pregnancy test results. The newer tests, which use monoclonal antibodies specific for the β subunit of HCG, also have increased analytical sensitivity. Radioimmunoassay (RIA) and enzyme-linked immunosorbent assay (ELISA) are able to detect 5 mIU/ml HCG in serum[39] contrasted to older polyclonal methods that could detect concentrations only as low as 100 mIU/ml. If a test is analytically nonspecific (it not only measures the target substances but also other closely related substances), the test will have low diagnostic specificity (incorrectly classifies disease negative individuals). Diagnostic tests with high analytical sensitivity and specificity do not necessarily produce high diagnostic sensitivity and specificity. Intervening variables such as spectrum bias, small sample aliquot and technical reliability of the assay can diminish diagnostic sensitivity and specificity.

Selection (Verification, Work-Up) Bias

Selection bias is a potential study design problem related to the way in which subjects are chosen for inclusion into a diagnostic test evaluation study. It is defined as the preferential referral of positive or negative test responders either to or not to verification diagnosis with a gold standard test.

Earlier in this lesson Riegelman and Hirsch's[6] basic steps to determine diagnostic performance listed step 2 as "perform the gold standard test on a full spectrum of subjects" and then step 3 as "test all subjects with the new diagnostic test." Under these procedures, the investigator starts with a group of individuals whose diagnosis has been verified with the gold standard. The investigator can select x number of verified disease positive and y number of verified disease negative subjects. The investigator thus determines the prevalence of the disease in the study's sample. Frequently an equal number of disease free and disease positive individuals are chosen as this provides the greatest statistical power for a given sample size.[6]

The important aspect of the above design is that all subjects receive the gold standard test. However, if the gold standard is an expensive test or if it is invasive and carries a high risk, another sample selection procedure might be used. Kelly, et al[40] present the example of a computed tomography (CT) scan compared to the gold standard of surgical inspection of the liver to diagnose liver lesions. Individuals 'who have a positive CT scan are referred for surgical resection of the liver. The verification diagnosis of liver pathology is made at the time of surgery. Individuals with negative CT scans are not referred for surgical verification. Work-up bias, in this example, will lead to high sensitivity and no meaningful value for specificity as there is no control group.

Avoidance of Work-up Bias Criteriion:[15]

- All subjects are assigned to receive both diagnostic and gold standard testing verification.

In some studies a partial solution is to refer a small random sample of the negative test subjects to the gold standard test for verification. Sensitivity and specificity calculations are done with modified equations that hope to correct the distortion caused by selection bias.[41] When the gold standard test is particularly risky or unethical to administer in perceived negative subjects, other mathematical corrections might be possible using retrospective adjustments with data from the source population.

Incorporation Bias

This bias occurs when the results of the diagnostic test being evaluated are incorporated into the gold standard testing procedure[22, 40] Incorporation bias would result if an initial diagnostic scan was done and then at a later time a second scan (the exact same procedure) was done to confirm the results of the first scan. The diagnostic test being evaluated and the gold standard test should be separate, independent procedures.

Diagnostic Review Bias

Diagnostic review bias occurs when the individual interpreting the results of the gold standard test know the results of the test being evaluated.[22,40] In this instance there is probably some subjectivity in the interpretation of the gold standard test results. Knowing the results of the diagnostic test can influence the care, scrutiny and objectivity of the gold standard test's interpretation. The solution is blinding the individual who evaluates the second test from the results of the first procedure.

Avoidance of Review Bias Criteriion:[15]
- Statement about independence in interpreting both the test and the gold standard procedure is included.

For an individual patient, if the results of the test being evaluated are true, then the carryover effect on the interpretation of the gold standard may not misrepresent the patient, but the comparison of results will be faulty. However, if the test being evaluated has misclassified the patient, carryover may misclassify the same patient in the same direction -the errors are correlated and the calculated sensitivity and specificity of the test being evaluated will be falsely high. Sensitivity and specificity are falsely high because the test being evaluated is misclassifying the same patient as the gold standard[20]

Test Review Bias

This is the opposite situation - the gold standard test results are known at the time the evaluated test results are reviewed.[22, 40] Again blinding is a powerful control to prevent bias in the interpretation of the test results.

Data Analysis

The analysis of the comparison of the diagnostic test to the gold standard test requires the independence of each test. The above biases illustrate violations of independence. If the test being evaluated and the gold standard test both misclassify the same patient, the tests have a falsely high agreement and sensitivity and specificity will.be falsely high. If the test being evaluated and the gold standard test independently misclassify the same patient, sensitivity and specificity will be underestimated.[20]

The diagnostic performance of a test is measured by the test's sensitivity, specificity, positive predictive value, negative predictive value, accuracy and likelihood ratios. The next section, medical decision making, will discuss each of these values. In addition to these diagnostic performance values, the investigators may also be interested in the relationship between the result values obtained from the

diagnostic tests they are comparing. Methods to analyze the relationship between two variables include regression analysis and correlation analysis.[42] If there are more than two variables, multiple regression analysis is used. If binary outcome variables are involved, multiple logistic regression and Cox regression methods are available. Regression analysis is a method of predicting one dependent variable from one or more independent variables; whereas, correlation analysis investigates whether or not there is a relationship between two variables and quantifies the relationship.

Correlation analysis can be useful in quantitating the relationship between the result values of two different diagnostic tests. The measurement scale of the diagnostic test result and the distribution (normal or not normal) characteristic of the values determines which analysis method is appropriate for the data. If both values are continuous scale and normal, Pearson correlation methods are used; if not normal, then rank correlation methods are appropriate. Rank correlation methods are also used if the result values are ordinal scale. When more than two variables are involved, multiple regression methods are used for continuous data and multiple logistic regression methods are used for binary data.

The relationship between the result values of two different diagnostic tests may be investigated even when the study will not support a full assessment of a new diagnostic test's accuracy. Flamen, et al.[43] compared rest and dipyridamole stress imaging using the myocardial perfusion agents, tetrofosmin and sestamibi. The investigators were interested in whether tetrofosmin might offer any advantage over sestamibi. The relationship between the segmental perfusion indices of tetrofosmin and sestamibi at rest and during stress was analyzed with linear correlation. A strong linear correlation was shown. However, because of the small sample size the investigators "did not attempt to study the absolute diagnostic accuracy of tetrofosmin."

Even when the full diagnostic test performance is assessed, the relationship between the test result values can be of interest. Inoue, et al.[44] compared two commonly used tumor seeking agents for PET [2-deoxy-2-^{18}F-fluoro-D-glucose (FDG) and L-methyl-^{11}Cmethionine (Met)] in detecting residual or recurrent malignant tumors in the same patients. The lesions were diagnosed based on pathological findings or clinical follow-up. The results showed similar, but limited sensitivity for FDG and Met (64.5% and 61.3% respectively) and significant correlation ($r = 0.788$, $p < 0.01$) between FDG and Met standardized uptake values (SUVs). The authors concluded the two PET agents were equally effective.

Indeterminate Test Results

For a variety of reasons the results of the test being evaluated may be indeterminate or uninterpretable. Examples from the literature include bowel gas obscuring the result of ultrasound, barium in the gastrointestinal tract obscuring the result of computed tomography, biopsy producing fragments that are insufficient for histological identification, and breast density invalidating mammography.[3]

Indeterminate test results should not be ignored. First the number of indeterminate test results that the diagnostic test generates is important. The patients with uninterpretable results will need further work-up. Either the test will have to be repeated or another test done. In either case, the patient will experience further expense and possible risk.

Presentation of Indeterminate Test Results Criteriion:[15]

- Report all of the appropriate positive, negative, and indeterminate results, and
- Report whether indeterminate results had been included or excluded when indexes of accuracy were calculated.

Within the diagnostic test evaluation study, if uninterpretable test results are counted as positive, sensitivity is falsely increased and specificity decreased. If the results are counted as negative, sensitivity is falsely decreased and specificity increased.[15] It is recommended that all indeterminate test results be reported as such by the investigators. Indeterminate results that happen as random events and with a test that is repeatable, can be disregarded in the analysis. If the indeterminate test results are related to the disease, it may be best to follow these patients or administer other diagnostic tests until their disease status is clear.

MEDICAL DECISION MAKING APPLIED TO DIAGNOSTIC TEST PERFORMANCE

We have just reviewed a litany of problems and biases that can interfere with a diagnostic test's performance. Important considerations were the kinds of patients included in the study and bias control in assessing the test results and the gold standard results. Methodologies to critically evaluate the diagnostic test's performance include: decision matrix, receiver operating characteristic (ROC) curve and information theory.[45]

Decision Matrix

Whether the diagnostic test results are measured in dichotomous scale or are measured in ordinal or continuous scale and converted to dichotomous scale, the results are presented for analysis in a two by two (fourfold) table, (see Table 1). The table logically relates the results of the diagnostic test to those of the gold standard test.

Sensitivity and Specificity

The binary results of the diagnostic test being evaluated are plotted on the two by two table, dividing the results between those that agree and those that do not agree with the gold standard test. Four cells are formed and thus four ratios generated that compare the results of the test to the actual presence or absence of disease.

The true positive (T+) ratio is the proportion of test positive (Test+) results among all disease positive [(D+)=(T+)+(F-)] individuals. This is the probability that patients with the disease will test positive. It is expressed as a conditional probability,

P(Test+ I D+),

the probability that a patient with the disease (D+) will have a positive test (Test+). The vertical bar is not a division symbol, but indicates the condition that is present or absent.[45] The true positive probability is the test's sensitivity. It expresses the test's accuracy in identifying patients with disease as positive.

The true negative (T-) ratio is the proportion of test negative (Test-) results among all disease negative [(D-)=(F+)+(T-)] individuals. This is the probability that patients without the disease will test negative. It is expressed as:

P(Test-1 D-),

the probability that a patient without the disease (D-) will have a negative test (Test-). The true negative probability is the test's specificity. It expresses that test's accuracy in identifying patients without the disease as negative Sensitivity and specificity represent the accuracy of the test itself. In the past these characteristics have been called the test's stable properties as they were thought to be stable over differing prevalences of the disease.

The overall accuracy of the test is:

$$\frac{(T+) + (T-)}{(T+) + (F+) + (F-) + (T-)}$$

Predictive Values

For an individual patient, sensitivity and specificity do not have personal meaning. At the time a diagnostic test is used on an individual patient to make decisions about treatment, the patients true disease status (gold standard test result) is unknown. For specific patients we would like to know how well the diagnostic test predicts their gold standard status. In other words, for patients testing positive, how likely is it that the patients are truly disease positive? If patients test negative, how likely is it that they are truly disease negative?

The horizontal features of the decision matrix provide the test's predictive values. The positive predictive value (PPV) of the test tells us what proportion of the test's positive results are truly positive. If the test is positive, how well does it "rule in" disease? The negative predictive value (NPV) of the test tells us what proportion of the test's negative results are truly negative. If the test is negative, how well does it "rule out" disease? At this point, please note that predictive values are proportions. Predictive values tell us what percentage of the test's positive (or negative) values are truly positive (or negative). The question we asked in the prior paragraph about how likely an individual patient's test result is true, is not really answered by predictive values. Likeliness is slightly different than proportionality and is discussed in the next section.

The overall accuracy of the test and the positive and negative predictive values do change with differing prevalences of the disease and are referred to as unstable. This is an important consideration for individual patients. The frequency with which diseases present themselves in various practice settings differs. For example, within a primary care setting we might expect to see a lower percentage of hypertensive patients than in a tertiary care setting. The predictive value of a positive or a negative test result will be different depending upon whether the patient was seen in a primary care clinic or a tertiary care setting.

Many early diagnostic test studies tend to use an artificial, 50:50, disease prevalence in the study. The predictive values in the study are thereby applicable only to populations where the disease prevalence is 50%. Most diseases do not occur that frequently. The results of the diagnostic test study can be translated to a local community if the local prevalence (albeit rough) of the disease is known or can be estimated. Using simple math, a two by two table can be constructed using the test's sensitivity and specificity and the local disease prevalence.[32] As the prevalence of most diseases is less than that used in diagnostic test evaluation studies, we would expect the positive predictive value of the test to

decrease and the negative predictive value to rise when the test's sensitivity and specificity values are applied to local prevalence. Table 2 illustrates this relationship. A diagnostic test is determined to have a sensitivity of 85% and a specificity of 90% in a research study that used a 50:50 sample prevalence. At 50% prevalence the predictive value of a positive test is 89.5% - a fairly reassuring value, indicating that 89.5% of patients with a positive test result are disease positive. The predictive value of a negative test is 85.7%. However, if the prevalence of the disease in the local community is only 1%, the predictive value of a positive test will be only about 8% - a much less satisfactory figure.

Likelihood Ratio

The primary indexes reported for diagnostic tests are sensitivity (true positive ratio) and specificity (true negative ratio). However, the corresponding false positive ratio and false negative ratio also have usefulness.

$$\text{False positive ratio} = P(\text{Test+} \mid \text{D-}) = (F+)/(F+)+(T-)$$

$$\text{False negative ratio} = P(\text{Test-} \mid \text{D+}) = (F-)/(T+)+(F-)$$

The false positive ratio is the proportion of positive tests in all patients who do not have the disease. The false negative ratio is the proportion of negative tests in all patients who have the disease. Ideally the diagnostic test would have high true positive and true negative ratios and also have low false positive and false negative ratios. The ability of a diagnostic test to identify individuals with disease, without incorrectly including patients without disease, that is, a high true positive and a low false positive ratio, is the likelihood ratio for a positive test result.[45]

$$\text{Likelihood Ratio for a Positive Test} = \text{true positive ratio}/\text{false positive ratio}$$

Tests with high likelihood ratios are desirable. The likelihood ratio for a positive test tells us how good the test will be at "ruling in" a disease. One can compare the likelihood ratios for two different diagnostic tests to determine which test does the better job of ruling in the target disease.[6]

In the same fashion, the ability of a diagnostic test to keep the number of disease positive individuals who are falsely identified as disease free to a minimum and also maximize the specificity of the test,

that is a low false negative and a high true negative ratio, is the likelihood ratio for a negative test result.

Likelihood Ratio for a Negative Test = false negative ratio/true negative ratio

For a likelihood ratio for a negative test, low values (less than 1) are desirable. The diagnostic test with the lowest likelihood ratio for a negative test result is the best test to "rule out" the target disease.

Likelihood ratios are actually risk ratios (or odds ratios). They are calculated by dividing the probability of a positive (or negative) test if the disease is present by the probability of a positive (or negative) test if the disease is absent. [6]

There are three characteristics of likelihood ratios that make them useful:[46]

- they are considered stable with changes in prevalence (although this premise has been challenged[36]),
- they can be calculated for every value of a continuous diagnostic test, instead of just two as sensitivity and specificity are, and
- used along with the patient's diagnostic test result, one can calculate the post-test probability that the patient has the target disease.

The last point is discussed further in the section, 'Post-Test Probability of Disease: Bayes Theorem.'

The size of the likelihood ratio is important, for both comparing which of two tests is the better for ruling in or ruling out a disease, and also for calculating the post-test probability of disease in a specific patient.[47] The following tabulation relates the actual size of a likelihood ratio to an ordinal scale description for example, a positive likelihood ratio with a value greater than 10 is considered large.

Table 3

CALCULATED LIKELIHOOD RATIOS EXAMPLE		
	Positive Likelihood Ratio	***Negative Likelihood Ratio***
Large	>10	<0.1
Moderate	5-10	0.1-0.2
Small, but sometimes important	2-5	0.5-0.2
Small, but rarely important	1-2	0.5-1.0

Table 3 contains an example of calculated likelihood ratios for various levels of a continuous scale diagnostic test taken from data presented by Vansteenkiste JF et al.[48] Patients with non-small-cell lung cancer (NSCLC) underwent thoracic computed tomography-(CT) scan, the radiolabeled glucose analog ^{18}F-fluoro-2-deoxy-D-glucose (FDG) positron emission tomography (PET) and invasive surgical staging (ISS). A five point visual scale was used for interpretation of lymph node (LN) staging on PET. Standardized uptake values (SUVs) were compared to the presence of metastasis in LNs and the likelihood ratios (LRs) for SUVs of LNs were determined. In this example the likelihood of metastasis in LNs increases as the SUV increases. For a SUV of LNs of >4.5 a positive test is 253 times more likely when metastasis is present than when metastasis is absent.

Receiver-Operating Characteristic (ROC) Curve

Ordinal or continuous scale diagnostic test results can be graphically represented. In Figure 2, if the cutpoint is decreased (moved left), the sensitivity (true positive probability) will be increased at the expense of specificity (false positives will increase). Sensitivity and specificity change as the cutpoint is moved lower or higher on the measurement scale. As measurements of diagnostic test accuracy, sensitivity and specificity do have some disadvantages. The indexes are probabilities and subject to manipulation. For example, if a disease is fairly uncommon and has a low prevalence, and if all patients who receive the diagnostic test are called negative, the indexes of accuracy would still be fairly

'accurate' as indeed most patients are negative. Even if decisions are not arbitrarily made, there is a certain amount of agreement between the diagnostic test being evaluated and the gold standard that can be expected by chance alone.

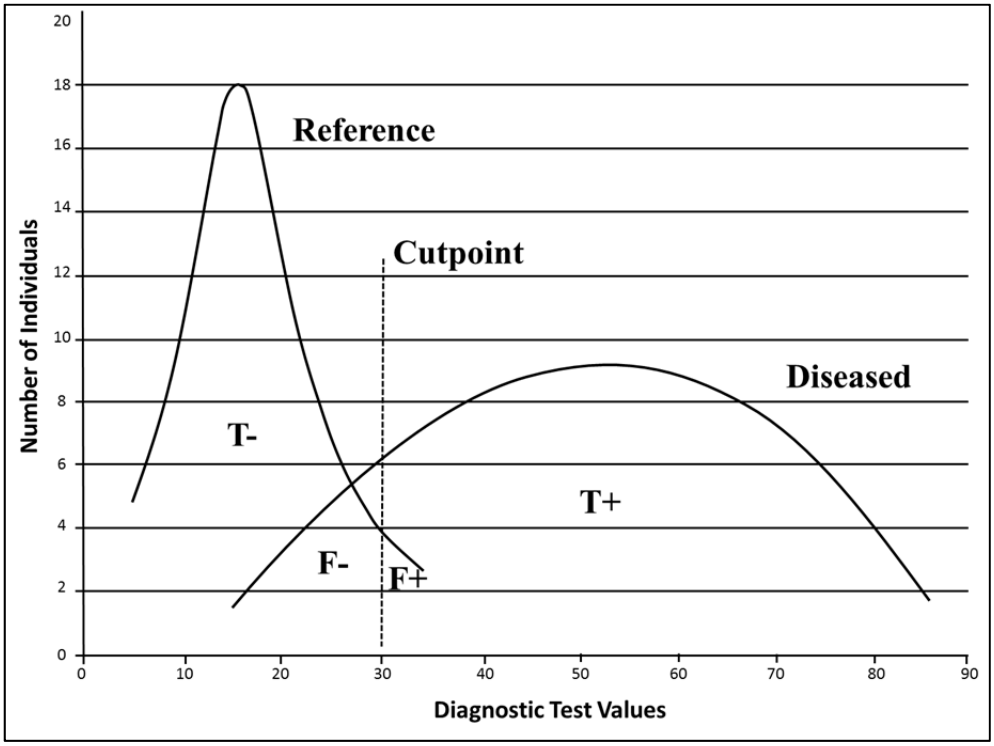

Figure 2. Characteristics of a Diagnostic Test

Metz[49] points out two major limitations of sensitivity and specificity. First, these indexes vary with disease prevalence. Secondly, using sensitivity and specificity to compare two different diagnostic tests is difficult as, while the percent of accurate results can be the same for two different tests, they can have different meanings. For example, one of the tests may error with false positive results and the other error with false negative results. Accuracy must always be considered in terms of both sensitivity and specificity. From Figure 2 we can see that the total number of disease free individuals (1 00% of reference sample) is made up of true negatives (T-) and false positive (F +) results. As fractions of the total then, (T- fraction) + (F + fraction) = 1. For the diseased sample, the total number of diseased individuals (100% diseased) is made up of true positive (T+) results and false negative (F-) results. Then, as fractions of the total, (T+ fraction)+ (F- fraction)= 1. Thus the value of each fraction will fall between 0 and 1. Using this terminology, the T + fraction is the sensitivity and the T- fraction is the specificity. Again, from Figure 2, one can see that if the reference and diseased curves overlap; the decision as to where on the scale of diagnostic test values the cutpoint is assigned is extremely important. Moving the cutpoint even a few units on the scale will change the number of individuals identified as falsely free of disease or falsely diseased. Regardless of how the cutpoint is determined, it is an arbitrary threshold between disease free and diseased.

Because the cutpoint is arbitrary, investigators often consider the accuracy indexes of sensitivity and specificity within a framework of several different cutpoints. A decision matrix can be constructed for each point on the measurement scale (x axis) that could be called either positive or negative - the values on the scale where the curves overlap. Each decision matrix will provide two sets of fractions, one the [T + fraction and F-fraction] which describes the diseased curve and the other the [T- fraction and the F+ fraction] which describes the reference curve. Since each set of fractions is equal to 1, if one of the fractions of a set is known, the other can be calculated. Likewise, if one fraction from each set is known, the characteristics of both sets are represented. By convention, the T + fraction (sensitivity) and the F + fraction (!-sensitivity) are chosen to represent the characteristics of accuracy for each cutpoint value that is considered. The T+ fraction and F+ fraction will increase or decrease between the value range of O to 1 as the cutpoint is moved up or down the diagnostic test measurement scale. The relationship of T+ fraction to F+ fraction over each of the 0 to 1 possible values can be graphically represented, Figure 3.

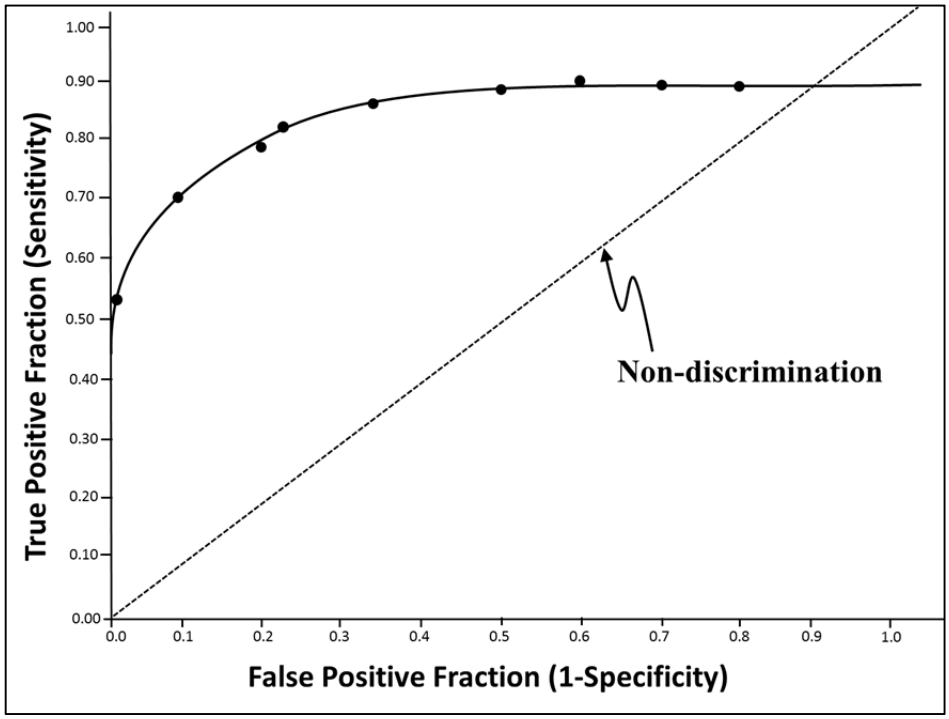

Figure 3. Receiver Operating Characteristic (ROC) Curve

The curve is called the receiver operating characteristic (ROC) curve, as it describes the characteristics of the test and the receiver of the test information can operate at any point on the curve.[49] Each point on the curve is a different cutoff point and is called an operating position.[45] If all test results can be called either positive or negative, the curve will pass through the (0,0) coordinate and through the (1,1) coordinate each point on the curve represents a different cutpoint and a different true positive to false

positive relationship. Points low on the curve represent a cutpoint having a low false positive, but also a low sensitivity. A point in the center of the curve provides a moderate sensitivity value for a moderate amount of false positive results. At the top of the curve, sensitivity is maximized at the expense of increased false positive results. Curves that are closest to the upper left margin provide the greatest sensitivity and the lowest false positive fraction. Curves that are further from the upper left quadrant and closer to the diagonal represent less desirable diagnostic test values. The diagonal line represents indices that would not discriminate as there would be an equal number of true positive results as false positive results.

The ROC curve is useful to visualize the consequences of choosing one cutpoint over others. It is also useful when comparing two or more competing diagnostic tests or two or more observers of diagnostic test results. Since true positive fraction and false positive fraction vary in direct relationship to one another, the true positive fractions (sensitivities) for each test being compared can be made to equal the same value. The false positive fraction for each test being compared would then adjust itself accordingly. The test with the lowest false positive fraction would be most desirable. Alternatively, the ROC curve for each test can be charted and the test with the largest area under the curve would provide the best sensitivity for the lowest false positive fraction. Metz[49] considered this the best method to compare differences in curves as there is not an acceptable mathematical method to test for statistically significant differences between ROC curves. The lack of an acceptable statistical test to show differences between ROC curves also means that there is no method of calculating a necessary sample size. From experience Metz [49] suggests a sample of about 100 or more patients is adequate. As with clinical trials, more subjects are need to find small differences between diagnostic test ROC curves than to find large differences.

ROC analysis of diagnostic test results requires the comparison of the test to a gold standard. However, as mentioned earlier, it is sometimes difficult to achieve a separate, independent gold standard comparison for diagnostic imaging tests as the gold standard is surgery or autopsy or includes using the results of the first imaging test with subsequent tests as the gold standard. There are now several reports in the literature of diagnostic test evaluation being done in the absence of a gold standard test. In the section on Gold Standard Choice we review such a study done by Line, et al.[19] A method to compare magnetic resonance imaging, computed tomography and radionuclide scintigraphy for the diagnosis of liver metastases without benefit of a gold standard has been presented by Henkelman, et al.[50]

According to Peirce and Cornell[51] the ROC originated in "World War II, when signal detection theory was applied to radar to *characterize* and evaluate how well a radar operator could receive or "see" a signal against a noisy background." In medicine, Lusted first applied ROC analysis to radiographic imaging in the late 1960's. The ROC curve is now the accepted method of summarizing ordinal and continuous scale diagnostic test accuracy data.

Information Theory

Information theory is concerned with reducing uncertainty. The primary function of a diagnostic test is to reduce the uncertainty of the patient's diagnosis. Stated another way, the diagnostic test should increase the amount of clinical information available to the medical decision-maker.

The decision of which of the contending cutpoints to choose to distinguish disease free from disease positive individuals can be addressed with information theory. The perfect diagnostic test would have a sensitivity of 100% (T + fraction = 1) and zero false positive results (F+ fraction= 0). At the other extreme, a non-discriminating test's ROC that offers a 50:50 chance of being positive or negative, would lie along the diagonal from 0,0 to 1, 1. Values below the major diagonal occur when positive results are more likely to occur when the test is negative than when it is positive. The perfect test provides maximum information content as it perfectly discriminates between disease positive and disease negative individuals. In practice diagnostic tests will not be perfect, but we hope they provide more information than that provided by a 50:50 chance of being positive or negative. Using the area under the curve model, if the perfect test has an area of 1; the non-discriminating diagonal has an area of0.5. The information content of each of the possible cutpoints on the ROC curve can be calculated.[52]

The choice of the optimal cutpoint also needs to be framed in the context of disease prevalence and values. Disease prevalence tells us how likely it is that any given patient could have the disease before the diagnostic test is done. If the disease is rare, we could guess that the patient does not have the disease and we would be correct most of the time. Under these circumstances, we may want to choose a cut point that is in the lower left portion of the curve where the sensitivity of the test is low, but the number of false positives will be lower still. If higher values on the curve are used, many of the positive individuals will be false positives. [45,49] Screening tests frequently fall into this situation. Also, if the consequences of a false positive decision are serious, such as surgery for an individual who does not have the disease, we will want to minimize false positives. If the prevalence of the disease is

higher, that is the disease is common, the best cutpoint will be in the upper right quadrant, where sensitivity is high along with high false positive results, but false negative results are minimized. If lower values on the curve are used, many of the negative results will be false negatives. This is also an appropriate strategy when it is extremely important to identify all individuals who have the disease.

The second consideration has to do with values. What is the value, cost, and consequence of assigning an individual to the false positive category or to the false negative category? Metz[49] provides mathematical methods to analyze the cost/benefit of each particular cutpoint. False negatives and false positives have both health costs and financial costs.

Lacking information about the health costs or financial costs of false positive and false negative results, a butpoint can be chosen that will minimize mistakes.[45] This is the point on the curve where the slope of the curve equals the prior probability of no disease, P(D), divided by the prior probability of disease, P(D+).

$$\text{Slope} = P(D-)/P(D+)$$

For example, if the pretest (prior) probability of disease in an individual patient is 15%, then the prior probability of no disease is 85%, and the best operating position on the curve is the point where the slope equals 0.85/0.15 = 5.7.

Post-Test Probability of Disease: Bayes Theorem

In the usual course of medical care the patient presents to the physician or the health care team. Using history and physical examination, the suspicion of disease in this patient develops into a possibility of disease. At this point a decision is made about whether or not diagnostic testing will occur. The function of the diagnostic test is to increase the information available to make decisions about the patient's health. If the diagnostic test cannot contribute further information to the scenario (the management of that patient is determined and wil1 not change), the test probably should not be done. In other words, we expect the diagnostic test to change our initial suspicion of disease in the patient - either the surety of no disease or the surety of disease should increase. In either case, before the diagnostic test is done, there is some suspicion of disease presence. The suspicion of disease presence can be expressed quantitatively as the probability that the patient has the specified disease. The diagnostic test results will, we hope, change the patient's probability of disease. How is the diagnostic test results combined with the pretest information about the patient to provide a post-test estimate of

disease presence? Bayes theorem can be used to revise the original estimate of disease in the patient by incorporating the information gained from the diagnostic test result. The words Bayes theorem and Bayesian inference are likely to cause our eyes to roll. However, what we are really concerned with is applying simple math to our natural decision making process. We develop a suspicion of disease in a patient and using history, physical examination and diagnostic tests (generally in series), we change/revise our suspicion as we gain more confidence from the new information. We use Bayes theorem to algebraically calculate the probability of disease after the positive (or negative) diagnostic test is known in light of the patient's initial (pretest) disease probability.

Our suspicion of disease after the diagnostic test, post-test probability of disease, is dependent upon the sensitivity and specificity of the diagnostic test and upon the pretest probability of disease in the patient. For an ideal (perfect) test the pretest probability of the disease is not important. If a test has perfect sensitivity, that is 100%, there are no false negative results (only true negative), therefore if the test is negative, the post-test probability of disease will be 0. If the test has perfect specificity, that is 100%, there are no false positives (only true positives), therefore if the test is positive, the post-test probability of disease will be 1. In this argument the post-test probability of disease is the same as the positive and negative predictive values.

Pretest Probability of Disease

Lacking a perfectly sensitive or specific test, the pretest probability of disease does influence the patient's post-test probability of disease. If the clinician estimates the patient's pretest probability of disease as low and the diagnostic test result is negative, the post-test probability of disease had changed little from the pretest value. However, if the same patient has a positive test result, the post-test probability of disease is quite different than the pretest value. The converse is also true, a high pretest probability of disease is changed little by a positive test result, but changed dramatically by a negative test result.

How do clinicians arrive at the pretest estimate of disease presence? Personal experience probably plays a large role in estimating pretest probabilities. The prevarence of the disease in the clinicians own practice area, such as the clinic or the hospital is most relevant. Other data sources may include community surveys or local databases. Such factors as the patient's age, sex, history and signs and symptoms will modify the local disease prevalence value. The pretest probability of disease is specific for each patient, so even local prevalence data will need to be personalized.

The published literature also provides estimates of prevalence for various diseases. The jump from published prevalence to a specific patient's pretest disease probability is large and presents problems. Published literature may be more representative of referral centers or tertiary care settings than the patient's local setting. Tertiary care setting will tend to have higher disease prevalence than primary care settings.

When faced with estimating the patient's pretest prevalence without good information, one method to somewhat control the uncertainty is to use the highest reasonable prevalence value and the lowest reasonable prevalence value and see if either of these values changes the patient's post-test probability of disease. Many clinicians intuitively assign a patient a pretest probability of disease, such as unlikely, likely, very likely. This inference model requires the ordinal estimate to be transformed to an explicit value.

One might ask about the disease prevalence that appears in the published diagnostic test evaluation study. Earlier we mentioned that in order to obtain the best precision in measuring all the diagnostic test's accuracy indices (T+, T-, F+, F-), investigators would enroll equal numbers of disease positive and disease negative subjects. While possible, it is not often that the 50% disease prevalence in the published study will match the patient's probability of disease. Thereby, the positive and negative predictive values from the published study will not apply to a specific patient.

There is also another reason to look further than the published study's positive and negative predictive values. In the examples presented so far and in Table 1, the likelihood ratio for a positive test and the likelihood ratio for a negative test were presented for just one cutpoint. For ordinal and for continuous data, we know that there is generally overlap between the disease negative and disease positive values. Because of this overlap, there is a 'sliding scale' of disease positivity. If only a single cutpoint is used to determine disease presence or absence, the information contained in the overlap area is lost. This information is retained if multiple cutpoints are used. This is done mathematically by using stratum-specific likelihood ratios.[51] In Table 3 standardized uptake values (SUV) in the locoregional lymph nodes (LNs) in patients with and without metastasis are divided into strata. For each stratum, the likelihood ratio is calculated. For an individual patient, the patient's diagnostic test result, in this example the SUV ofLNs, can be matched to one of the stratum. If the patient's SUV was 4.0, the 3.5-4.5 stratum applies and the corresponding likelihood ratio (LR) is 3.157.

Earlier we discussed ROC curves, a plot of (sensitivity) vs. (1 -specificity); it is interesting to note that a likelihood ratio for a positive test is (sensitivity)/(1 - specificity).

Where do we find likelihood ratios? Despite the apparent sensibleness of likelihood ratios, particularly spectrum-specific likelihood ratios, they have not been overwhelming adopted into clinical practice. However, their use is increasing and they have been strongly promoted by evidence based medicine publications. Likelihood ratios are being reported in published diagnostic test evaluation studies. List of likelihood ratios have been generated by McMaster University[46] and also appear in newer textbooks.[53]

Calculating Post-test Probability

> "…diagnostic tests simply change the odds in favor or against disease; sometimes they only confuse the diagnosis."[54]

Stratum-specific likelihood ratios are considered a better way of reflecting diagnostic test accuracy than sensitivity and specificity for a single cutpoint.[46, 54] Given this, likelihood ratio is the index used to incorporate the patient's pretest disease probability into the diagnostic test result to produce the post-test probability. Referring to Table 1 we can see that the likelihood ratio for a positive (or negative) test result is a ratio of two probabilities: the probability of a positive (or negative) test if the disease is present to the probability of a positive (or negative) test if the disease is absent. Likelihood rations are actually risk ratios or odds rations. The fact that the desirable index to use to calculate post-test probability of disease is an odds ratio is unfortunate. Most health care professionals think in terms of probabilities, not odds. In fact, we have just spoke of the pretest and post-test probabilities of disease for a specific patient – not pre- and post-test odds. To combine a likelihood ration with the pretest probability of disease, we must convert the pretest probability of disease to odds. The formula is:[51, 55]

Pretest probability/(1-pretest probability) = pretest odds

If the probability of disease is 15%, then the disease odds is equal to (0.15/(1-0.15) = (0.15/0.85) = 0.18:1.

The post-test odds of disease is calculated by:

Post-test odds = pretest odds x likelihood ratio

The odds ratio form of Bayes' theorem requires that the pretest probability of disease be converted to odds and that the resultant post-test odds of disease be converted back to a probability.

Post-test probability = post-test odds/(post-test odds+ 1)

The McMaster University group46 has published a nomogram, Figure 4, that relieves us of the calculations. The nomogram appears on their web site (see site 2 in the following Internet Educational Resources section). An interactive nomogram is available from site 3 in the Internet Educational Resources listing. The nomogram is used by drawing a straight line from the pretest probability value through the likelihood ratio for the diagnostic test and continuing the line to the post-test probability scale.

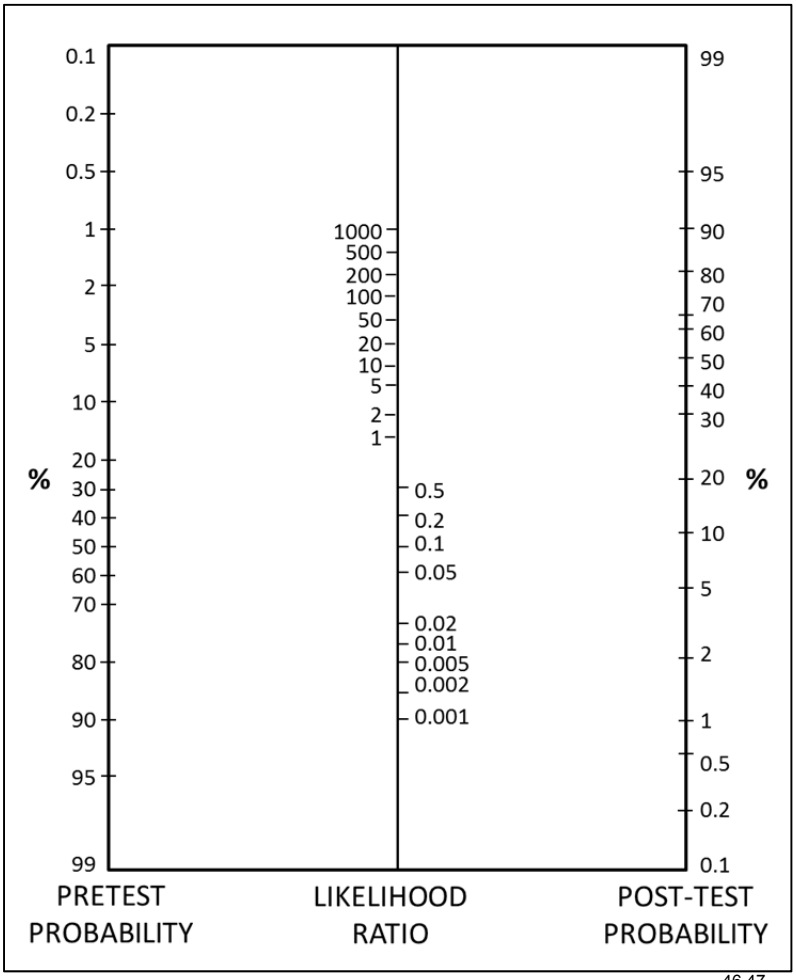

Figure 4. Nomogram for Interpreting Diagnostic Test Results[46,47]

A research publication by K wok, et al. [56] illustrates the use of the nomogram. The K wok article is discussed further in the meta-analysis section of this lesson. The study generated summary diagnostic test indices for exercise electrocardiogram (ECG), exercise thallium, and exercise echocardiogram (echo) for the diagnosis of coronary artery disease in women. The indices are:

	Sensitivity	Specificity	Likelihood Ratio (+)	Likelihood Ratio (-)
ECG	0.61	0.70	2.25	0.55
Thallium	0.78	0.64	2.87	0.36
Echo	0.86	0.79	4.29	0.48

The authors concluded that these exercise tests were only moderately sensitive and specific for diagnosing coronary artery disease (CAD) in women. The likelihood ratios fall within the small to moderate value range. The authors went on to calculate the posttest probabilities of CAD for both a positive and a negative test result for women with three different pretest CAD probabilities. For a woman with nonspecific chest pain (pretest probability of CAD of 6%), using echo testing (the test with the best values) the post-test probability of a positive test is 21% and 1% for a negative test. The post-test diagnosis is less clear for women with an intermediate pretest probability of 31 %; the post-test probability of CAD for a positive test is 66% and 7% for a negative test result. Values very close to those calculated can be obtained by using the pretest probabilities and the likelihood ratio with the nomogram (Figure 4). The pretest probability of CAD for a woman with definite angina is 71%.

DIAGNOSTIC TESTS IN PRACTICE

Screening and Case Finding

Diagnostic tests that are done to find undiagnosed disease in otherwise healthy individuals are called screening or case finding tests. Some authors make a distinction between the two. ‘Screening’ is used when the health care provider initiates the encounter. Testing for hypertension in a community screening program is an example. 'Case finding' is differentiated as the patient initiates the encounter. The patient seeks medical care and during this encounter, diagnostic tests are done to detect 'silent' disease processes. Regardless of who initiates the encounter, the diagnostic test is done because the disease presents a significant health concern, not because the patient has a complaint related to the target disease.

Examples of screening tests include routine testing for fasting blood sugar to detect diabetes mellitus, occult blood in stool to detect rectal cancer and mammography to detect breast cancer. To find all cases of a disease, the sensitivity of the screening test should be fairly high. As further testing and evaluation will be done in positive cases, the test does not necessarily need high specificity. J

Diagnosis

Contrasted to screening and case finding procedures, diagnosis proceeds because there is some suspicion that the patient has a disease condition. The patient has initiated the encounter or been referred to the provider because of symptoms possibly due to a disease. The health care provider would like at this point a diagnostic test that will either confirm or exclude the suspected disease.

An individual diagnostic test may be very good for screening, confirmation or exclusion - some tests can even do two of the three very well. However, very few tests can appropriately be used for screening, confirmation and exclusion. Even if a test could reasonably do all three, expense and risk to patient may limit the test's usefulness.

Confirmatory Tests

If there is a strong suspicion of disease, a confirmatory test will be used to verify the diagnosis. For example, if a patient has had a positive fasting blood glucose, a glucose tolerance test might be used to confirm the diagnosis.

To confirm a diagnosis, a diagnostic test that is capable of ruling in the diagnosis is needed. Contrary to what we might first think, a diagnostic test with a high specificity works well to rule in a diagnosis. A positive test result from a diagnostic test that has a high true negative and low false positive proportion will effectively rule in a diagnosis. The low false positive proportion also gives the test a high positive predictive value. SpPin (high *Specificity*, *Positive* result rules *in* the diagnosis) describes this situation.[51] For example, a histologically positive bronchoscopic biopsy will rule in lung cancer, but a negative biopsy does not rule out the disease.

Not all diagnostic tests will have a conveniently high specificity. Since the patient's post-test probability of having the disease is dependent upon not only the results of the diagnostic test, but also the pretest disease probability, we might also look at the test's likelihood ratio. Likelihood ratios (LR) of 10 or higher for a positive test result can significantly increase the patient's probability of disease. Even if the patient's pretest probability of disease is only 10%, a LR of 10 can increase the post-test probability of disease to about 53%.

Tests with LR of 5 to less than 10 are intermediate, causing moderate shifts in the pretest probability of disease. Compared with diagnostic tests having only one reported LR, those with stratum-specific likelihood ratios provide more precise estimates of the individual patient's post-test disease probability.

Exclusionary Tests

An exclusionary test provides the evidence necessary to rule out a disease in a patient who has some suspicion of disease. If the patient has a negative test result from a diagnostic test with a high sensitivity (low false negative proportion and high negative predictive value), we can be fairly confident the patient does not have the disease. A highly *s*ensitive test with a *n*egative result will rule *out* a disease, SnNout. The skin test for tuberculosis is an example of a test whose negative results rul~ out the disease, but a positive test result does not rule in tuberculosis. The test is also a good screening test.

A likelihood ratio for a negative test result of less than 0.1 (low false negatives) can significantly change the patient's pretest disease probability away from a positive diagnosis. LRs of0.2 to 0.1 will have a moderate influence on the patient's pretest disease probability.

Diagnostic Tests in Combination

Diagnostic tests are of value when the suspicion of disease is high enough to stimulate some action, but not yet high enough to be a foregone conclusion with no further need for diagnostic procedures. A single diagnostic test is unlikely to screen, confirm and exclude patients for a given disease. At the point of initial suspicion the clinician has to decide if a single test will be done first (with the option to then move on to other tests – series approach) or if several tests will be done in parallel. In either strategy, a single diagnostic test is rarely sufficient. Each test that is done should bring incremental value to the diagnosis.

In the series strategy, since only the positive patients will receive the second test, the first test should be a screening test with high sensitivity (SnNout). The screening test will filter out the true negative patients (and, hopefully only a small number of false negatives). The remaining positive test result patients include both true and false positive individuals. These patients will require a second confirmatory test or group of tests that will definitively verify the presence of disease. The series strategy will not work if the disease is likely to change quickly, before the second test or set of tests is

performed. There should not be any carryover or influence of the first test on the performance of the second test(s).

The parallel testing strategy may work well if there is no single test with high sensitivity or high specificity. For example, there may be two tests with only modest or low sensitivity or specificity, but each test picks up on a different type of disease, such as one detects early disease and the other one detects later symptoms. Using the two tests in combination will identify four groups: both tests are positive (disease present), both tests are negative (disease absent) and one test is positive and the other negative and the converse (further testing required).

When multiple diagnostic tests are used, a new source of bias enters the arena. For continuous scale diagnostic test results we know that there is generally overlap between the true negative and the true positive test values. Commonly a central percentage, such as 95%, of the negative values are called negative and the balance, such as 5%, are called 'abnormal.' If only a single diagnostic test is done, the percentage of misdiagnosed individuals will be 5%. If multiple diagnostic tests are performed this percentage of misdiagnosed individuals increases.[7] For two tests, the central percentage of disease negative changes from 95% to $(.95)^2 = 90\%$, the balance, 10% will be misdiagnosed. For a twenty test chemistry profile, the central disease negative proportion drops from 95% for a single test to $(.95)^{20} = 36\%$; the probability of a false positive diagnosis with one of the chemistries is now 64%. Multiple diagnostic tests present the same problem as multiple statistical significance testing, each statistical test done increases the probability of a false positive result. Clinicians may want to consider wider boundaries for the disease negative values to reduce the chance of false positive results.

Meta-analysis of Diagnostic Tests

Meta-analysis is not the combining of different diagnostic tests in the same patient, but the combining, mathematically, of the results of different research studies evaluating the same diagnostic test. The purpose of meta-analysis is to achieve a stronger point estimate of the target outcome measured by combining like studies. It is a useful technique when several studies have reported equivocal results or when all the studies have relatively small sample sizes. Meta-analysis is a separate study design and we will not review the design in this lesson. However, meta-analyses of diagnostic tests are becoming more common. Kwok, et al[56] reviewed published studies that evaluated the accuracy of exercise electrocardiogram (ECG), exercise thallium, and exercise echocardiogram (echo) for the diagnosis of coronary artery disease in women. Using standard meta-analysis methods, the quality of each study

was judged by (1) the inclusion of an adequate description of the persons selected for the study, (2) the absence of verification bias, and (3) absence of review bias if the results were read blindly. The authors constructed a decision matrix and ROC curve for each study. A summary ROC curve for the studies weighted by sample size was then plotted. By using summary data for each test, the authors were able to draw conclusions on the value of all three tests for the diagnosis of coronary artery disease in women.

Table 1. Decision Matrix

		Disease Present (D+)	Disease Absent (D-)	
Test Results	Disease Evident (Test +)	True Positives (T+) Likelihood (T+) / (T+)+(F-)	False Positives (F+) Likelihood (F+) / (F+)+(T-)	**Positive Predictive Value** (T+) / (T+)+(F+)
	Disease Not Evident (Test -)	False Negatives (F-) Likelihood (F-) / (T+)+(F-)	True Negatives (T-) Likelihood (T-) / (F+)+(T-)	**Negative Predictive Value** (T-) / (F-)+(T-)
		Sensitivity (T+) / (T+)+(F-)	**Specificity** (T-) / (F+)+(T-)	**Prevalence** (T+)+(F-) / (T+)+(F+)+(F-)+(T-)
		Likelihood Ratio Test+ (T+)/(T+)+(F-) / (F+)/(F+)+(T-)	**Likelihood Ratio Test-** (F-)/(T+)+(F-) / (T-)/(F+)+(T-)	**Accuracy** (T+)+(T-) / (T+)+(F+)+(F-)+(T-)

Table 2. Predictive Values of Test Results for Various Disease Prevalences

Disease Prevalence (Pretest Probability) Sensitivity 85%, Specificity 90%				
	1%	***5%***	***10%***	***50%***
Predictive Value of a Positive Test	7.9%	30.9%	48.6%	89.5%
Predicative Value of a Negative Test	99.8%	99.1%	98.2%	85.7%

Table 3

Likelihood Radios (Probability of Metastasis) for a Continuous Scale Diagnostic Test (Standardized Uptake Value of Lymph Nodes) [89]			
SUV of LNs	Lymph Node Staging for Metastasis		
	Metastasis Present	Metastasis Absent	
	Likelihood	Likelihood	Likelihood Ratio
<3.5	*0.149*	*0.6977*	*0.152*
3.5 4.5	0.064	0.020	3.157
>4.5	0.787	0.003	253.096

Data is based on invasive surgical staging of 690 lymph node stations.

GLOSSARY OF DIAGNOSTIC TEST INDEX TERMS			
Test Characteristic	***Synonym***	***Explanation***	***Formula***
Sensitivity	P (Test + | D+) True positive proportion or ratio	The ability of the test to reliably detect disease	(T+) I (T+) + (F-)
Specificity	P (Test + | D-) True negative proportion or ratio	The ability of the test to reliably detect the absence of disease	(T-) I (F+) + (T-)
Positive Predictive Value	Post-test probability of a positive test	Proportion of patients with a negative test that actually are disease free	(T+) I (T+) + (F+)
Negative Predicative Value	Post-test probability of a negative test	The probability of having the disease at a point in time	(T-) I (F-) + (T-)
Prevalence	Pretest Probability of Disease Prior Probability of Disease	Overall agreement of test with the gold standard	(T+) + (F-) I (T+) + (F+) + (F-) + (T-)
Accuracy		The odds that a given positive diagnostic test result would be expected in a patient with (as opposed to one without) the target disease[40]	(T+) + (T-) I (T+) + (F+) + (F-) + (T-)
Likelihood Ratio for a Positive Test Result	Ratio of (T + ratio) to the (F+ratio)	The odds that a given positive diagnostic test results would be expected in a patient with (as opposed to one without) the target disease.	Sensitivity I (1 -specificity) or [(T +) I (T +) + (F-)] I [(F+) I (F+) + (T-)]
Likelihood Radio for a Negative Test Result	Ratio of (F – ratio) to the (T-ratio)	The odds that a given negative diagnostic test result would be expected in a patient with (as opposed to one without) the target disease.	(1 -sensitivity) I specificity or [(F-) I (T+)+(F-)] I [(T-) I (F+)+(T-)]]

REFERENCES

1. WulffHA. Rational diagnosis and treatment. Oxford: Blackwell Scientific Publications, 1976:78.

2. Fineberg HV, Bauman R, Sosman M. Computerized cranial tomography, effect on diagnostic and therapeutic plans. JAMA 1977;238:244-7.

3. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;6:411-23.

4. Mackenzie R, Dixon AK. Measuring the effects of imaging: an evaluative framework. Clin Radiol 1995;50:513-8.

5. Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. AJR 1994;162:1-8.

6. Riegelman RK and Hirsch RP. Studying a study and testing a test, how to read the health science literature. 3rd ed. Boston: Little, Brown and Company; 1996.

7. Earl RA. Establishing and using clinical laboratory reference ranges. Appl Clin Trials 1997;6:24-30.

8. Kerlinger FN. Foundations of behavioral research, 3nd ed. New York: Holt, Rinehart and Winston, Inc., 1986.

9. Anon. How to read clinical journals: II. To learn about a diagnostic test. CMA Journal1981; 124:703-10.

10. Campbell DT and Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally College Publishing Company; 1963.

11. Rosner B. Fundamental of biostatistics, 3rd edition. Boston: PWS-Kent Publishing Company, 1990.

12. Birnbaum D, Sheps SB. Validation of new tests. Infect Control Hosp Epidemiol1991;12:622-4.

13. Koran L. The reliability of clinical methods, data and judgments. N Engl J Med 1975;293:695-701.

14. Anderson RE, Hill RB, Key CR. The sensitivity and specificity of clinical diagnostics during five decades, toward an understanding of necessary fallibility. JAMA. 1989; 261:1610-7.

15. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research, getting better but still not good. JAMA. 1995; 274:645-51.

16. Jaeschke R, Guyatt G, Sackett DL: Users' guides to the medical literature Ill. how to use an article about a diagnostic test, A. are the results of the study valid? JAMA. 1994; 271:389-91.

17. Arroll B, Schechter MT, Sheps SB. The assessment of diagnostic tests: a comparison of medical literature in 1982 and 1985. J Gen Intern Med. 1988; 3:443-7.

18. Hubbard WK. Regulations for in vivo radiopharmaceuticals used for diagnosis and monitoring, Federal Register 1999(May 17);64:26657-70.

19. Line BR, Peters TL, Keenan J. Diagnostic test comparisons in patients with deep venous thrombosis. J Nucl Med 1997;38:89-92.

20. Valenstein PN. Evaluating diagnostic tests with imperfect standards. Am J Clin Pathol. 1990; 93:252-8.

21. Dewey ME. Designs for studies evaluating tests. Intematl J Geriatric Psychiatry. 1997; 12:492-4.

22. RansohoffDF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978; 299:926-30.

23. Morise AP, Diamond GA, Detrano R, Bobbio M. Incremental value of exercise electrocardiography and thallium-20 1 testing in men and women for the presence and extent of coronary artery disease. Am Heart J. 1995; 130:267-76.

24. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? JAMA 1990; 263:275-8.

25. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence, sample size estimation for diagnostic test studies. J Clin Epidemiol1991; 44(8):763-70.

26. Linnet K. Comparison of quantitative diagnostic tests, type I error, power, and sample size. Statistics in Medicine 1987;6: 147-58.

27. Buderer NMF. Statistical methodology, I. incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med 1996;3:895-900.

28. Freedman LS. Evaluating and comparing imaging techniques, a review and classification of study designs. Br J Radiology 1987;60:1071-81.

29. Kent DL, Larson EB. Health policy in radiology, disease, level of impact, and quality of research methods, three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Investigative Radiology 1992; 27:245-54.

30. Morton RF, Hebel JR, McCarter RJ. A study guide to epidemiology and biostatisics, 3rd ed. Gaithersburg, MD: Aspen Publishers, Inc.; 1989.

31. Coughlin SS, Pickle LW. Sensitivity and specificity-like measures of the validity of a diagnostic test that are corrected for chance agreement. Epidemiology 1992;3: 178-81.

32. Anon. How to read clinical journals: II. to learn about a diagnostic test. Can Med Assoc J 1981;124: 703-10.

33. Greenhalgh T. How to read a paper, papers that report diagnostic or screening tests. BMJ 1997;315: 540-3.

34. Leisenring W, Pepe MS, Longton G. A marginal regression modelling framework for evaluating medical diagnostic tests. Stat in Med 1997;16: 1263-81.

35. Vecchio TJ. Predictive value of a single diagnostic test in unselected populations. N Engl J Med 1966; 274:1171-3.

36. Brenner H, Gefeller 0. Variation of sensitivity, specificity, likelihood ratios and predictive values with' disease prevalence. Stat in Med 1997;16:981-91.

37. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Reports 1947;62:1432-49.

38. Sailh AJ, Hoover DR. Lingua medica, "sensitivity" and "specificity" reconsidered, the meaning of these terms in analytical and diagnostic settings. Ann Intern Med 1997;126:91-4.

39. McCombs J, Cramer MK. In: Pharmacotherapy, a pathopysiologic approach. 4th ed. DePiro JT, Talbert RL, Yee GC, et al., eds. Stamford, Connecticut: Appleton and Lange: 1999:1298.

40. Kelly S, Berry E, Roderick P, Harris KM, Cullingworth J, Gathercole L, Hutton J, Smith MA. The identification ofbias in studies of the diagnostic performance of imaging modalities. Br j Radiology 1997; 70:1028-35.

41. Choi BCK. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. J Clin Epidemioll992;45:58l-6.

42. Rosner B. Fundamentals ofbiostatistics, 3rd ed. Boston: PWS-Kent publishing Company; 1990.

43. Flamen P, Bossuyt A, Franken PR. Technetium-99m-tetrofosmin in dipyridamole-stress myocardial SPECT imaging, intraindividual comparison with technetium-99m-sestamibi. J Nucl Med 1995;36:2009-15.

44. Inoue T, Kim EE, Wong FCL, et al. Comparison offluorine-18-fluorodeoxyglucose and carbon-11-methionine PET in detection of malignant tumors. J Nuc Med 1996;37:1472-6.

45. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl JMed 1975;293:211-5.

46. Anon. Interpretation of diagnostic data, 5. How to do it with simple math. Can Med Assoc J 1983; 129:947-54.

47. Jaeschke R, Guyatt GH, Sackett DL. How to use an article about a diagnostic test. [resource on World Wide Web]. URL: http://hiru/hirunet.mcmaster.ca/ebm/userguid/3 dx.htrn. Accessed 1999 Mar 28. This information is also available as: Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature, III. How to use an article about a diagnostic test, B. what are the results and will they help me in caring for my patients? JAMA 1994;271 :703-707. .

48. Vansteenkiste JF, Stroobants SG, DeLeyn PR, et al. Lymph node staging in non-small-cell lung cancer with FDG-PET scan, a prospective study on 690 lymph node stations from 68 patients. J

Clin Oncol 1998; 16:2142-9.

49. Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Medicine 1978;8(4):283-98.

50. Henkelman RM, Kay I, Bronskill MJ. Receiver operator characteristic (ROC) analysis without truth. Med Decis Making 1990;10:24-9.

51. Peirce JC and Cornell RG. Integrating stratum-specific likelihood ratios with the analysis of ROC curves. Med Decis Making 1993;13:141-51.

52. Metz CE, Goodenough DJ, Rossmann K. Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. Radiology 1973;109:297-303.

53. Sloane PD, Slatt LM, Curtis P, Ebel, eds. Essentials of family medicine, 3rd ed. Baltimore: Williams and Wilkins; 1998.

54. Sackett DL, Straus S. On some clinically useful measures of the accuracy of diagnostic tests. ACP Journal Club 1998;129:A-17. Simel DL. Playing the odds. Lancet 1985(Feb 9);1:329. Kwok Y, Kim C, Grady D, Segal M, Redberg R. Meta-analysis of exercise testing to detect coronary artery disease in women. Am J Cardioll999;83:660-6.

ASSESSMENT QUESTIONS

Read the following abstract and construct a 2 by 2 table for scintimammography compared to the gold standard, biopsy:

Title: Mammography and 99mtc-mibi scintimammography in suspected breast cancer
Author(s): Prats E; Aisa F; Abos MD; Villavieja L; Et Al
Source: J Nucl Med, vol40, iss 2, p 296-301, yr 1999

Article Abstract: The aim of this work has been to evaluate whether a diagnostic protocol based on the joint use of mammography and 99mTc-methoxyisobutyl isonitrile (MIBI) scintimammography is capable of reducing the number of biopsies required in patients with suspected breast cancer.

Methods: We performed prone scintimammography in 90 patients with suspected breast cancer, involving 97 lesions. In all patients, the diagnosis was established by way of biopsy. On mammography, we evaluated the degree of suspicion of malignancy and the size of the lesion (smaller or larger than 1 cm in diameter).

Results: The results of only 41 of the biopsies indicated malignancy. On mammography, 20 lesions (of which 1 was breast cancer) were considered to be of low suspicion of malignancy, 31 (of which 4 were breast cancer) as indeterminate and 46 (of which 36 were breast cancer) as high. Fourteen lesions (2low probability, 2 indeterminate and 10 high) were smaller than 1 cm, whereas 83 (18 low probability, 29 indeterminate and 36 high) were larger. Scintimammography results were positive in 35 cases of breast cancer. Scintimammography was positive in all cases of breast cancer that initially had a low or indeterminate suspicion of malignancy according to mammography, as well as in 30 cases of breast cancer that initially were highly suspicious. Six false-negative scintimammography studies were obtained. In the benign lesions, scintimammography results were positive in 12 cases and normal in 44.

Conclusion: We propose a diagnostic protocol with a biopsy performed on lesions that have a high suspicion of malignancy as well as those with low or indeterminate suspicion that are smaller than 1 cm or with positive scintimammography results. This would have reduced the total number of biopsies performed by 34%. More importantly, there would have been a 65% reduction in number of biopsies performed in the low and indeterminate mammographic suspicion groups. All 41 cases of breast cancer would have been detected.

Sensitivity = proportion of those with the disease who are correctly identified by the test [True Positive Ratio (TPR)]

Specificity = proportion of those without the disease who are correctly identified by the test [True Negative Ratio (TNR)]

Predictive value of a positive test = proportion of those with a positive test who have the disease

Predictive value of a negative test= proportion of those with a negative test who do not have the disease

1. Compute the sensitivity (TPR) of scintimammography to detect malignant lesions:

    a. 6%
    b. 15%
    c. 36%
    d. 79%
    e. 85%

2. Computer the specificity (TNR) scintimammography to detect malignant lesions:

    a. 6%
    b. 12%
    c. 14%
    d. 79%
    e. 85%

3. Compute the predictive value of a negative scintimammography:

    a. 12%
    b. 25%
    c. 74%
    d. 88%
    e. 96%

4. Compute the predictive value of a negative scintimammography:

    a. 6%
    b. 12%
    c. 45%
    d. 52%
    e. 88%

5. Compute the overall accuracy of scintimammography:

    a. 36%
    b. 45%
    c. 52%
    d. 81%
    e. 100%

6. Within this study, what is the prevalence of malignant breast lesions?

    a. 6%
    b. 36%
    c. 42%
    d. 73%
    e. 85%

7. The sensitivity and specificity of a diagnostic test provide the clinical data to support which of the following hierarchical levels used to associate a diagnostic test procedure with a patient's health outcome?

    a. Technical performance of the test [reliability]
    b. Diagnostic performance [accuracy]
    c. Diagnostic impact [displaces alternative tests]
    d. Therapeutic impact [influence on treatment tests]
    e. Impact on health [quality of life]

Your screening facility can process 1,000 people per week. Assume you are attempting the

8. Individual who are healthy and truly do not have the disease condition under consideration belong to the group of individuals who are best described as:

    a. Control group
    b. Intervention group
    c. True negative group
    d. Normal group
    e. Reference group

9. As a part of a routine physical examination, uric acid was measured for a 35-year-old male and found to be 7.8 mg/dl. The "normal range" for uric acid for that laboratory is 3.4 to 7.5 mg/dl. If this individual does not display symptoms or signs of gout, a possible explanation is:

    a. He is among the small proportion of healthy individuals who yield high serum uric acid readings on a given test.
    b. His level is within 2 standard deviations of the mean for healthy individuals
    c. His test results represent a false negative.
    d. The departure of his level from the normal range is statistically significant.
    e. This individual is a good candidate for early treatment

10. A pulmonary angiogram is a highly sensitive test (considered the gold standard) for pulmonary embolus. For a patients who has several general symptoms (shortness of breath, vague chest pain) consistent with pulmonary embolus, a **negative** test result:

    a. Implies that the disease is less prevalent in this patient's population.
    b. Indicates that the patient has only a mild embolus.
    c. Will require the patient's close blood relatives to be tested.
    d. Will rule in the disease.
    e. Will rule out the disease.

A 35 year-old female complains of a mild burning pain on urination and some lower abdominal

11. As part of the quality control procedures for a diagnostic laboratory, great care is usually invested to ensure that those individuals who calibrate the equipment, execute the test and record the data all follow the same procedures for each step of the diagnostic test. The laboratory is seeking to avoid which one of the following factors that can jeopardize the validity of the test?

    a. History
    b. Maturation
    c. Testing
    d. Instrumentation
    e. Statistical regression

12. Which of the following statements about bias is FALSE?

    a. Good laboratory procedures can eliminate some bias
    b. Bias is systemic error
    c. The presence of bias decreases the internal validity of a diagnostic test study
    d. The presence of bias decreases the external validity of a diagnostic test study
    e. Bias is random error

13. A study, that compares one diagnostic test to second diagnostic test that is thought to be the gold standard for diagnosis of the disease under consideration, is seeking to establish which of the following kinds of validity for the first test?

    a. Content validity
    b. Criterion-related validity
    c. Construct validity

14. The technical performance of a new diagnostic test is established by the test's:

    a. Incremental cost
    b. Complexity of execution
    c. Reliability
    d. Content validity
    e. Manufacturer

15. In a trial investigation risk factor for breast cancer, women received yearly mammograms. The trial was multicentered and over the years, one center had reported significantly fewer positive mammograms. An investigation pointed to a lack of experience and training on the part of those investigators reading mammograms at that site. This problem is one of:

    a. Selection bias
    b. Fuzzy trial hypothesis
    c. Generalizability
    d. Validity
    e. Reliability

16. To achieve the status of being the gold standard for diagnosis of a given disease conditions, a diagnostic test must fit which of the following?

    a. Be generally accepted as the best available diagnostic test
    b. Accurately diagnose the disease status of every patient
    c. Be capable of being executed (used) in both ambulatory and inpatient settings
    d. Have the best ROC curve
    e. Be the least invasive diagnostic test

17. In practice, the presence or absence of disease in an individual patient is generally accepted if the diagnosis was established using any of the following methods, **EXCEPT:**

    a. The test with the smallest false negative fraction
    b. Definitive histophathologic diagnosis
    c. Standard diagnostic classification system
    d. Well-established diagnostic tests
    e. Patient follow-up

18. Blood glucose values are on a continuous scale and by changing the cut point for being positive or negative for diabetes, one can change the sensitivity and specificity of the test. If sensitivity and specificity for several cut points on the scale were calculated, then a receiver-operating-characteristic (ROC) curve could be drawn. Which one of the following is **FALSE**

    a. The ROC curve could be used to choose the best cutoff point to define an abnormal blood glucose level depending up on emphasis placed on health costs, financial costs or information content of the test
    b. An ROC curve strategy would work particularly well for a binary response diagnostic test
    c. To construct a ROC curve sensitivity (TPR) and 1 – specificity (FNR) constitute the vertical and horizontal axis of the ROC graph respectively
    d. A diagonal line from 0, 0 to 1, 1 represents indices that do not discriminate between true positive results and false positive results.
    e. A lax threshold for a positive diagnosis can be described as highly sensitive, but having poor specificity.

19. **Title:** The diagnostic accuracy of bedside and laboratory coagulation, procedures used to monitor the anticoagulation status of patients treated with heparin

Article Abstract: We evaluated the diagnostic accuracy of three bedside coagulation procedures, the Hemochron, activated whole-blood clotting time (ACT), the CoaguChek Plus, activated partial thromboplastin time (APTT) and theTAS, APTT, in patients who received heparin therapy. As part of the patients' care, pharmacists performed bedside coagulation tests. Blood from heparinized patients was analyzed with each of the three tests and a gold standard laboratory test. Receiver operator characteristic (ROC) curves were plotted for each test. Analysis of the ROC curve was used to rank the performance of the methods. Areas under the ROC curves+/- SE for the CoaguChek Plus APTT, Hemochron ACT, and TAS APTT were 0.872 +/- 0.044, 0.797 +/- 0.039, and 0.795 +/- 0.048, respectively.

The laboratory test that demonstrated the highest diagnostic accuracy (maximizes the true positives and minimizes the false positives) for predicting who is and who is not anticoagulated by heparin is which of these three tests?

a. CoaguCheck Plus
b. Hemochron
c. TAS
d. They are all equivalently accurate
e. None of the three have AUC's less than .45, thus none of them have diagnostic discrimination.

20. The ability of a diagnostic test study to accurately determine the sensitivity and specificity of the diagnostic test is dependent upon how many individuals participate in the study (sample size). The equation to calculate sample size contains each of the following elements **EXCEPT:**

a. Alpha level (Type I error)
b. Beta level (Type II error)
c. Number of investigators
d. Variability of events
e. Delta (clinically meaningful values for sensitivity and specificity)

21. Power is a statistical term used to define the probability of detecting a meaningful difference between two treatments when there is one. "Meaningful difference" is determined by:

a. P values of ≤ 0.05
b. Values falling outside 2 standard deviations from the mean
c. The investigator's judgment
d. Standards established in the clinical guidelines for each disease condition
e. Type II error rate

22. Investigators faced with inadequate subject accrual into a clinical trial may decide to continue the study with a smaller sample size. The consequence of this action on the study's validity is to:

    a. Alter the subjects' maturation characteristics
    b. Increase cost
    c. Decrease the study's power
    d. Jeopardize voluntary consent
    e. Unbalance baseline characteristics

23. The disease prevalence within the diagnostic study's subject that will achieve the greatest statistical power for the study is:

    a. 25%
    b. 50%
    c. 75%
    d. 96%
    e. 100%

24. Subjects within a diagnostic test study that should receive the gold standard diagnostic test include:

    a. All subjects
    b. 50% of the male and 50% of the female subjects
    c. Subjects whose test results fall into the true positive range of values
    d. All subjects who finish the study
    e. Subjects whose test results are equivocal

25. Whether or not sensitivity and specificity are independent of disease prevalence has been controversial. However, for diagnostic situations that are strongly dichotomous, sensitivity and specificity are considered independent of prevalence. Which of the following conditions provides dichotomous test results?

    a. Asthma
    b. Coronary artery disease
    c. Hypertension
    d. Parkinson's disease
    e. Pregnancy

26. Diagnostic tests frequently not only diagnose the presence or absence of disease, but also measure an endogenous substance in the patient's body. Diagnostic tests that have relatively poor analytic specificity will also have:

    a. Poor diagnostic specificity
    b. Good diagnostic specificity
    c. Usage only in analytical applications
    d. Poor diagnostic sensitivity
    e. Usage in diagnostic applications

27. Diagnostic review bias tends to produce falsely high sensitivity and specificity values. The design method used to avoid this bias is:

    a. Randomly select study subjects
    b. Randomly allocate study subjects to new the new test and the gold standard test
    c. Blind the individual who interprets the results of the new test and the gold standard test
    d. Use confidence intervals
    e. Calculate the study's power

28. Meta-analysis is a useful technique to combine the results of different studies that investigated the same diagnostic test procedure. Which of the following statements does **NOT** describe a situation where meta-analysis could be used?

    a. Study data are in disagreement as to the magnitude of the accuracy indices
    b. Studies all used the same independent variable, i.e., the diagnostic test procedure
    c. Sample sizes in individual studies are too small to reliably detect diagnostic accuracy
    d. Large (sample size) trials are not feasible
    e. Only data from case series reports are available

29. Clin Nucl Med, vol 20, iss 9, p 821-829, yr 1995 "Tc-99m sestamibi demonstrates considerable renal uptake followed by net urinary clearance similar to that of creatinine. The authors have previously shown that renograms could be obtained in cardiac patients by imaging during the rest injection of the perfusion agent. The present study shows correlating Tc099m sestamibi and Tc-99m DTPA studies in hypertensive patients with a spectrum of findings, includes aortic aneurysms, asymmetry due to renovascular disease, cysts, bilateral renal dysfunction, and horseshoe kidney."

    Which of the following statements is **TRUE**?

    a. Spectrum bias can be decreased by using hypertensive patients with a spectrum of renal disorders
    b. Spectrum bias decreases the study's internal validity
    c. Excluding patients with comorbidities will control spectrum bias
    d. Studies with a narrow spectrum of subjects tend to underestimate the test's accuracy
    e. Spectrum bias has only been shown to operate on sex and race demographic variables.

30. The predictive value of a positive lung imaging diagnostic test will vary depending upon whether or not the patient comes from the general ambulatory care population or from a tertiary care veterans institution population.

    a. This statement is true
    b. This statement is false

31. The predictive value of a negative diagnostic test is **highest** when:

    a. The patient has no other diseases
    b. There is no accepted gold standard for the disease
    c. The disease prevalence is very low
    d. The test is performed at a large medical center
    e. The test agrees with the physician's opinion

32. If a patient is suspected of having the disease in question, to confirm the diagnosis one would challenge the patient with a diagnostic test that:

    a. Has a low negative likelihood ratio
    b. Has a low positive likelihood ratio
    c. Has a high sensitivity
    d. Has a high positive likelihood ratio
    e. Has a high cost

33. The best source of data to use for the pretest probability of a disease is:

    a. Textbooks
    b. Published current literature
    c. The World Wide Web
    d. Tertiary care center data
    e. Local community data

34. Exclusionary tests are used to rule out the target disease in a patient who has some suspicion of disease. An exclusionary test should have:

    a. High sensitivity
    b. High specificity
    c. Low sensitivity
    d. Low specificity
    e. Moderate accuracy

35. The primary function of a diagnostic test is:

    a. Generate a positive cash flow for the department
    b. Corroborate the opinion of the physician
    c. Satisfy the patient
    d. Reduce uncertainty
    e. Establish a patient database