Chapter 14
Research Methodology for Studies of Diagnostic Tests

Stephen P. Glasser

Research is what I'm doing when I don't know what I'm doing.
Wernher von Braun http://www.brainyquote.com/quotes/authors/w/wernher_von_braun.html
Prediction is very difficult, especially about the future.
http://larry.denenberg.com/predictions.html

Abstract Much of clinical research is aimed at assessing causality. However, clinical research can also address the value of new medical tests, which will ultimately be used for screening for risk factors, to diagnose a disease, or to assess prognosis. In order to be able to construct research questions and designs involving these concepts, one must have a working knowledge of this field. In other words, although traditional clinical research designs can be used to assess some of these questions, most of the studies assessing the value of diagnostic testing are more akin to descriptive observational designs, but with the twist that these designs are not aimed to assess causality, but are rather aimed at determining whether a diagnostic test will be useful in clinical practice. This chapter will introduce the various ways of assessing the accuracy of diagnostic tests, which will include discussions of sensitivity, specificity, predictive value, likelihood ratio, and receiver operator characteristic curves.

Keywords Predictive value • Sensitivity • Specificity • Receiver operator curves • Bayes' Theorem • Likelihood ratio • Net reclassification index • Test accuracy

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_14,

Introduction

Up to this point in the book, we have been discussing clinical research predominantly from the standpoint of causality. Clinical research can also address the value of new medical tests, which will ultimately be used for screening for risk factors, to diagnose a disease, or to assess prognosis. The types of research questions one might formulate for this type of research include: "How does one know how good a test is in giving you the answers that you seek?" or "What are the rules of evidence against which new tests should be judged?" In order to be able to construct research questions and designs involving these concepts, one must have a working knowledge of this field. Although traditional clinical research designs can be used to assess some of these questions, most of the studies assessing the value of diagnostic testing are more akin to descriptive observational designs, but with the twist that these designs are not aimed to assess causality, but are rather aimed at determining whether a diagnostic test will be useful in clinical practice.

Bayes Theorem

Thomas Bayes was an English theologian and mathematician who lived from 1702 to 1761. In an essay published posthumously in 1863 (by Richard Price), Bayes' offers a solution to the problem "…to find the chance of probability of its happening (a disease in the current context) should be somewhere between any two named degrees of probability" [1]. Bayes' Theorem provides a way to apply quantitative reasoning to the scientific method. That is, if a hypothesis predicts that something should occur and it does, it strengthens our belief in that hypothesis; and, conversely if it does not occur, it weakens our belief. Since most predictions involve probabilities i.e. a hypothesis predicts that an outcome has a certain % chance of occurring, this approach has also been referred to as probabilistic reasoning. Bayes' Theorem is a way of calculating the degree of belief one has about a hypothesis. Said in another way, the degree of belief in an uncertain event is conditional on a body of knowledge (this is in contrast to the traditional statistical model called the frequentist approach which does not incorporate prior knowledge in its statistical calculations). Suppose we're screening people for a disease (D) with a test that gives either a positive or a negative result (A and B, or T+ and T− respectively). Suppose further that the test is quite accurate, in the sense that, for example, it will give a positive result 95 % of the time when the disease is present (D+), i.e. $P(T+|D+) = 0.95$ (this formula asks what is the probability of the disease being present GIVEN a positive test?), or said another way, what is the probability that a person who tests positive has disease? The naive answer is 95 %; but this is wrong. What we really want to know clinically is $P(D+|T+)$, that is, what is the probability of testing positive if one has the disease; and, Bayes' theorem (or predictive value) tells us that.

In modern medicine the first useful application of Bayes' theorem was reported in 1959 [2]. Ledley and Lusted demonstrated a method to determine the likelihood that a patient had a given disease when various combinations of symptoms known to be associated with that disease were present [2]. Redwood et al. utilized Bayesian logic to reconcile seemingly discordant results of treadmill exercise testing and coronary angiography [3]. In 1977, Rifkin and Hood pioneered the routine application of Bayesian probability in the non-invasive detection of coronary artery disease (CAD) [4]. This was followed by other investigative uses of Bayesian analysis, an approach which has now become one of the common ways of evaluating all diagnostic testing.

As noted above, diagnostic data can be sought for a number of reasons beside just the presence or absence of disease. For example, the interest may be the severity of the disease, the ability to predict the clinical course of a disease, or to predict a therapy response. For a test to be clinically meaningful one has to determine how the test results will affect clinical decisions, what are its cost, risks, and what is the acceptability of the test; in other words, how much more likely will one be about this patients problem after a test has been performed than one was before the test; and, is it worth the risk and the cost? Recall, that the goal of studies of diagnostic testing seeks to determine whether a test is useful in clinical practice. To derive the latter we need to determine whether the test is reproducible, how accurate it is, whether the test affects clinical decisions, etc. One way to statistically assess test reproducibility (i.e. inter and intra-variability of test interpretation), is with a kappa statistic [5]. Note that reproducibility does not require a gold standard, while accuracy does. In order to talk intelligently about diagnostic testing, some basic definitions and understanding of some concepts is necessary.

Kappa Statistic (k)

The kappa coefficient is a statistical measure of inter-rater reliability. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Cohen's kappa measures the agreement between two raters [5].

The equation for κ is:

$$\Pr(a) - \Pr(e) / 1 - \Pr(e)$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the probability that agreement is due to chance.

If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$ (See Table 14.1). Note that Cohen's kappa measures agreement between two raters only. For a similar measure of agreement when there are more than two raters Fleiss' kappa is used [5]. An example of the use of the kappa statistic is shown in Table 14.2.

Table 14.1 Strength of agreement using the kappa statistic

Kappa	Strength of agreement
0.00	Poor
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Table 14.2 An example of the use of the kappa statistic

		Doctor A		
		No	Yes	Total
Doctor B	No	10(34.5 %)	7(24.1 %)	17(58.6 %)
	Yes	0(0.0 %)	12(41.4 %)	12(41.4 %)
Total		10(34.5 %)	19(65.5 %)	29

Kappa = (Observed agreement – Chance agreement)/(1 – Chance agreement)
Observed agreement = (10 + 12)/29 = 0.76
Chance agreement = 0.586 * 0.345 + 0.655 * 0.414 = 0.474
Kappa = (0.76 – 0.474)/(1 – 0.474) = 0.54

Definitions

Pre-test Probability

The pre-test probability (likelihood) that a disease of interest is present or not, is the index of suspicion for a diagnosis, ***before*** the test of interest is performed. This index of suspicion is influenced by the prevalence of the disease in the population of patients you are evaluating. Intuitively, one can reason that with a rare disease (low prevalence) that even with a high index of suspicion, you are more apt to be incorrect regarding the disease's presence, than if you had the same index of suspicion in a population with high disease prevalence.

Post-test Probability and Test Ascertainment

The post-test probability is one's index of suspicion ***after*** the test of interest has been performed. Let's further explore this issue as follows. If we construct a 2 × 2 table (Table 14.3) we can define the following variables: If disease is present and the test is positive, that test is called a true positive (TP) test (this forms the definition of test sensitivity – that is the % of TP tests in patients with the index disease). If the index disease is present and the test is negative, that is called a false negative (FN) test. Thus, patients with the index disease can have a TP or FN result (but by definition cannot have a false positive – FP, or a true negative -TN result).

Table 14.3 The relationship between disease and test result

	Abnormal test	Normal test
Disease present	True positive (TP)	False negative (FN)
Disease absent	False positive (FP)	True negative (TN)

Table 14.4 An example of the pre and post-test probability given disease prevalence and the sensitivity and specificity of a test

Pre vs post-test probability		
Prev = 10 % of 100 patients, Se = 70 %, Sp = 90 %		
	T+	**T−**
D+	7/10 (TP)	3/10 (FN)
D−	9/90 (FP)	81/90(TN)
	PV+ 7/16 = 44 % (10 % → 44 %)	
	PV− 81/84 = 97 % (90 % → 96 %)	

Sensitivity and Specificity

The sensitivity of a test then can be written as TP/TP+FN. If the index disease is not present (i.e. it is absent) and the test is negative, this is called a true negative (TN) test (this forming the definition of specificity-that is the % of TN's in the absence of disease). The specificity of a test can then be written as TN/TN+FP. Finally, if disease is absent and the test is positive one has a false positive (FP) test. Note that the FP % is 1-specificity (that is, if the specificity is 90 % – in 100 patients without the index disease, 90 will have a negative test, which means 10 will have a positive test – i.e. FP is 10 %).

Predictive Value

Another concept is that of the predictive value (PV+ and PV−) of a test. This is asking the question differently than what sensitivity and specificity address – that is rather than asking what the TP and TN rate of a test is, the PV+ of a test result is asking how likely is it that a positive test is a true positive (TP)? i.e. TP/TP+FP (for PV− it is TN/TN+FN). See the example of the calculation of PV in Table 14.4.

Ways of Determining Test Accuracy and/or Clinical Usefulness

There are at least six ways of determining test accuracy and they are all interrelated so the determination of which to use is based on the question being asked, and one's personal preference. They are:

Sensitivity and Specificity
2 × 2 Tables

Predictive Value
Bayes Formula of Conditional Probability
Likelihood Ratio
Receiver Operator Characteristic Curve (ROC)

Bayes Theorem

We have already discussed sensitivity and specificity as well as the tests predictive value, and the use of 2 × 2 tables; and, examples will be provided at the end of this chapter. But, understanding Bayes Theorem of conditional probabilities will help provide the student interested in this area with greater understanding of the concepts involved. First let's discuss some definitions and probabilistic lingo along with some shorthand. The conditional probability that event A occurs given population B is written as P(A|B). If we continue this shorthand, sensitivity can be written as P(T+|D+) and PV+ as P(D+|T+). Bayes' Formula can be written then as follows: The post test probability of disease =

$$\frac{(\text{Sensitivity})(\text{disease prevalence})}{(\text{Sensitivity})(\text{disease prevalence})+(1-\text{specificity})(\text{disease absence})}$$

or

$$\frac{P(D+|T+)=P(T+|D+)(\text{prevalence } D+)}{P(T+|D+)(\text{prevalence } D+)P(T+|D-)P(D-)}$$

where P(D+|T+) is the probability of disease given a T+ (otherwise known as PV+), P(T+|D+) is the shorthand for sensitivity, P(T+|D−) is the FP rate or 1-specificity. Some axioms apply. For example, one can arbitrarily adjust the "cut-point" separating a positive from a negative test and thereby change the sensitivity and specificity. However, any adjustment that increases sensitivity (this then increases ones comfort that they will not "miss" any one with disease as the false negative rate necessarily falls) will decrease specificity (that is the FP rate will increase – recall 1-specificity is the FP rate). An example of this is using the degree of ST segment depression during an electrocardiographic exercise test that one has determined will identify whether the test will be called "positive" or "negative". The standard for calling the ST segment response as positive is 1 mm of ST segment depression from baseline, and in the example in Table 14.2 this yields a sensitivity of 62 % and specificity of 89 %. Note what happens when one changes the definition of what a positive test is, by using 0.5 mm ST depression as the cut-point for calling test positive or negative. Another important axiom is that the prevalence of disease in the population you are studying does not significantly influence the sensitivity or specificity of a test (to derive those variables the denominators are defined as subjects with or without the

Table 14.5 An example of calculating post test probability of disease using Bayes formula

Pre vs post-test probability		
Prev = 50 % in 100 patients, Se = 70 %, Sp = 90 %		
	T+	**T−**
D+	.7 × 50 = 35 (TP)	.3 × 50 = 15 (FN)
D−	.1 × 50 = 5 (FP)	.9 × 50 = 45 (TN)
	PV+ 35/40 = 87 %	
	PV− 45/60 = 75 %	

$$P(D+T+) = \frac{.7(.5)}{.7(.5)+1-.9(.5)} = \frac{.35}{.35+.05} = .87$$

Table 14.6 Estimations of pre and post test probabilities of disease given the clinical presentation

Pre vs post-test probabilities			
Clinical presentation	Pre test P (%)	Post test P T+ (%)	Post test P T− (%)
Typical angina	90	98	75
Atypical angina	50	88	25
No symptoms	10	44	4

disease i.e. if you are studying a population with a 10 % disease prevalence one is determining the sensitivity of a test – against a gold standard- only in those 10 %). In contrast, PV is very dependent on disease prevalence because more individuals will have a FP test in populations with a disease prevalence of 10 % than they would if the disease prevalence was 90 %. Consider the example in Tables 14.5 and 14.6.

Receiver Operator Characteristic Curves (ROC)

The ROC is another way of expressing the relationship between sensitivity and specificity (actually 1-specificity). It plots the TP rate (sensitivity) against the FP rate over a range of "cut-point" values (actually the ROC curve is a plot of likelihood ratios – see below). It thus provides visual information on the "trade off" between sensitivity and specificity, and the area under the curve (AUC) of a ROC curve is a measure of overall test accuracy (Fig. 14.1). ROC analysis was born during WW II as a way of analyzing the accuracy of sonar detection of submarines and differentiating signals from noise [6]. In Fig. 14.2, a theoretic "hit" means a submarine was correctly identified, and a false alarm means that a noise was incorrectly identified as a submarine and so on. You should recognize this figure as the equivalent of the table above discussing false and true positives.

Another way to visualize the tradeoff of sensitivity and specificity and how ROC curves are constructed is to consider the distribution of test results in a population. In Fig. 14.3, the vertical line describes the threshold chosen for a test to be called positive or negative (in this example the right hand curve is the distribution of

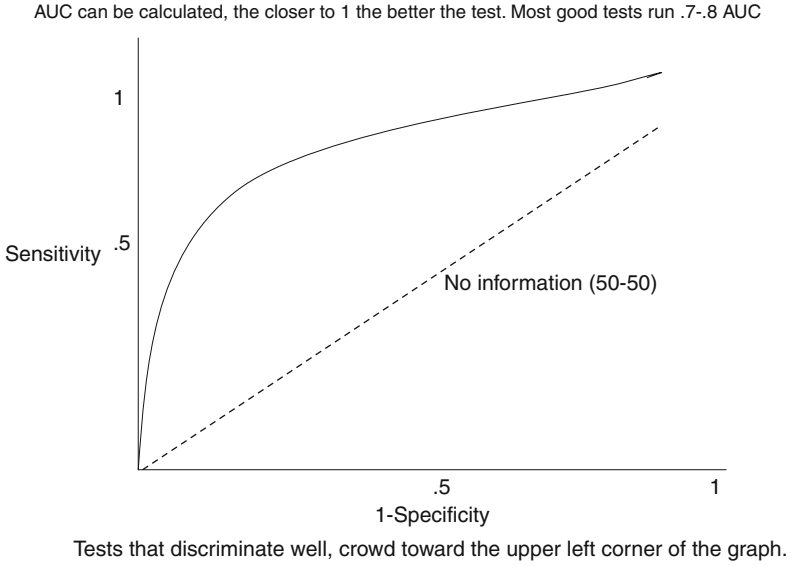

Fig. 14.1 An example of a Receiver Operator Characteristic (ROC) curve

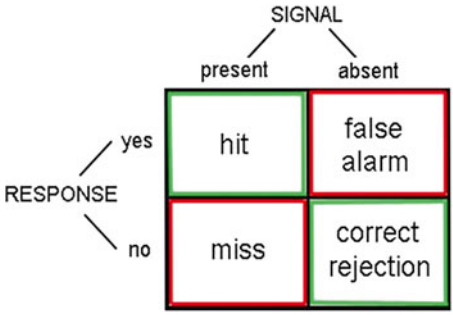

Fig. 14.2 A diagram of the use of sonar to correctly identify submarines (http://www-psych.stanford.edu/~lera/psych115s/notes/signal/. Accessed 11/05/2013)

subjects within the population that have the disease, the left hand curve those who do not have the disease). The uppermost figure is an example of choosing a very low threshold value for separating positive from negative. By so doing, very few of the subjects with disease (recall the right hand curve) will be missed by this test (i.e. the sensitivity is high-97.5 %), but notice that 84 % of the subjects without disease will also be classified as having a positive test (false alarm or false+rate is 84 % and the specificity of the test for this threshold value is 16 %). By moving the vertical line (threshold value) we can construct different sensitivity to false+rates and construct a ROC curve as demonstrated in Fig. 14.4.

As mentioned before, ROC curves also allow for an analysis of test accuracy (a combination of TP and TN), by calculating the area under the curve as shown in the figure above. Test accuracy can also be calculated by dividing the TP and TN by all

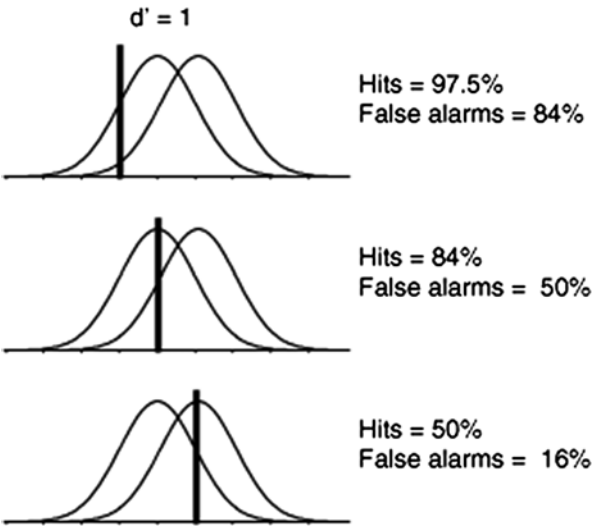

Fig. 14.3 An example of how moving the definition of positive vs negative tests alter the results of correctly identifying a target (http://www-psych.stanford.edu/~lera/psych115s/notes/signal/. Accessed 11/05/2013)

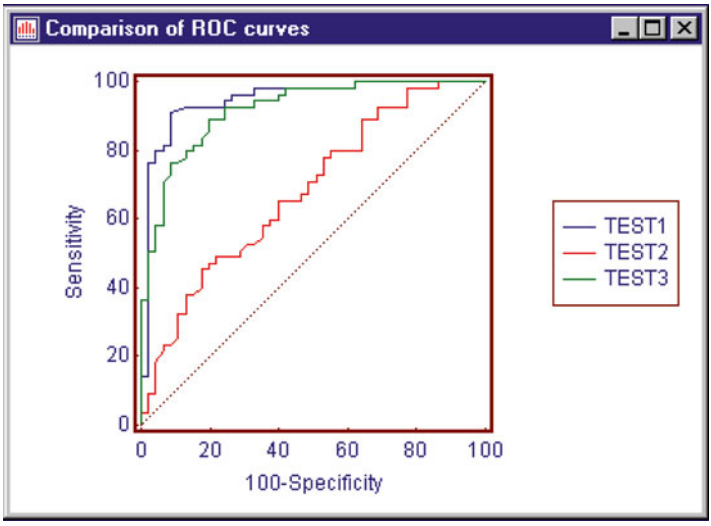

Fig. 14.4 Examples of ROC curves from three different tests (http://en.wikipedia.org/wiki/Receiver_operating_characteristic. Accessed 11/05/2013)

possible test responses (i.e. TP, TN, FP, FN). The way ROC curves can be used during the research of a new test, is to compare the new test to existent tests as demonstrated by Maisel et al. [7].

Likelihood Ratios

Positive and Negative Likelihood Ratios (PLR and NLR or LR+ and LR−) are another way of analyzing the results of diagnostic tests. Essentially, PLR is the odds that a person with a disease would have a particular test result, divided by the odds that a person without disease would have that result. In other words, how much more likely is a test result to occur in a person with disease than a person without disease. If one multiplies the pretest odds of having a disease by the PLR, one obtains the posttest odds of having that disease. The PLR for a test is calculated as the tests sensitivity/1-specificity (i.e. FP rate). So a test with a sensitivity of 70 % and a specificity of 90 % has a PLR of 7 (70/1 − 90). Unfortunately, it is made a bit more complicated by the fact that we generally want to convert odds to probabilities. That is, the PLR of 7 is really an odds of 7 to 1 and that is more difficult to interpret than a probability (the probability from a 7:1 odds is 87.5 %, see below). Recall that odds of an event are calculated as the number of events occurring, divided by the number of an events *not* occurring (i.e. non events, or p/p − 1). So if blood type O occurs in 42 % of people, the odds of someone having a blood type of O are .42/1 − .42 i.e. the odds of a randomly chosen person having blood type O is .72:1. Probability is calculated as the odds/odds + 1, so in the example above .72/1.72 = 42 % (or .42 – that is one can say the odds have having blood type O is .72 to 1 or the probability is 42 %-the latter is easier to understand for most). Recall, that probability is the extent to which something is likely to happen. To review, take an event that has a 4 in 5 probability of occurring (i.e. 80 % or .8). The odds of its occurring is 0.8/1 − 0.8 or 4:1. Odds then, are a ratio of probabilities. Note that an odds ratio (often used in the analysis of clinical trials) is also a ratio of odds.

To review:

The likelihood ratio of a positive test (LR+) is usually expressed as

$$\text{Sensitivity} / 1 - \text{Specificity}$$

and the LR− is ***usually*** expressed as

$$1 - \text{Sensitivity} / \text{Specificity}$$

If one has estimated a pretest odds of disease, one can multiply that odds by the LR to obtain the post test odds, i.e.:

$$\text{Post - test odds} = \text{pre - test odds} \times \text{LR}$$

To use an exercise test example consider the sensitivity for the presence of CAD (by coronary angiography) based on 1 mm ST segment depression. In this aforementioned example, let's assume that the sensitivity of a "positive" test is 70 % and the specificity is 90 % (PLR = 7; NLR = .33). Let's assume that based upon our

Table 14.7 Different ways of calculating Likelihood Ration (LR)

End point	LR	Ratio	Se:Sp
D+ for T+	**LR+**	**%D+ with T+**	**Se/1–SP**
		%D– with T+	**TP/FP**
D– for T–	LR–	%D– with T–	Sp/1–Se
		%D+ with T–	TN/FN
D– for T+	1/LR+	%D– with T+	1–Sp/Se
		%D+ with T+	FP/TP
D+ for T–	**1/LR–**	**%D+ with T–**	**1–Se/Sp**
		%D– with T–	**FN/TN**

history and physical exam we feel the chance of a patient having CAD before the exercise test is 80 % (0.8). If the exercise test demonstrated 1 mm ST segment depression, your post-test odds of CAD would be .8 × 7 or 5.6 (to 1). The probability of that patient having CAD is then 5.6/1 + 5.6 = .85 (85 %). Conversely if the exercise test did not demonstrate 1 mm ST segment depression the odds that the patient did not have CAD is .33 × 7 = 2.3 (to 1) and the probability of his not having CAD is 70 %. In other words *before* the exercise test there was an 80 % chance of CAD, while *after* a positive test it was 85 %. Likewise before the test, the chance of the patient *not* having CAD was 20 %, and if the test was negative it was 70 %.

To add a bit to the confusion about using LRs, there are two lesser-used derivations of the LR as shown in Table 14.7. One can usually assume that if not otherwise designated, the descriptions for PLR and NLR above apply. But, if one wanted to express the results of a negative test in terms of the chance that the patient **has** CAD (despite a negative test) rather than the chance that he **does not** have disease given a negative test; or wanted to match the NLR with NPV (i.e. the likelihood that the patient does NOT have the disease given a negative test result) an alternative definition of NLR can be used (of course one could just as easily subtract 70 % form 100 % to get that answer as well). To make things easier, a nomogram can be used instead of having to do the calculations [8].

In summary, the usefulness of diagnostic data depends on making an accurate diagnosis based upon the use of diagnostic tests, whether the tests are radiologic, laboratory based, or physiologic. The questions to be considered by this approach include: "How does one know how good a test is in giving you the answers that you seek?", and "What are the rules of evidence against which new tests should be judged?" Diagnostic data can be sought for a number of reasons including: diagnosis, disease severity, to predict the clinical course of a disease, to predict therapy response. That is, what is the probability my patient has disease x, what do my history, physical exam, and baseline laboratory data tell me, what is my threshold for action, and how much will the available tests help me in patient management. An example of the use of diagnostic research is provided by Miller and Shaw, which demonstrates how the coronary artery calcium (CAC) score can be stratified by age and the use of the various definitions described above [9].

Beyond the ROC Curve

Over 30 years after the construction of the first multivariable risk prediction model predicting the probability of developing cardiovascular disease (CVD) new risk factors that can predict CVD and that can be incorporated into risk assessment algorithms has progressed. An individual's age, baseline levels of systolic and diastolic blood pressure and serum cholesterol, smoking and diabetes status are all useful predictors of the CVD risk over a reasonable future time period, typically 1–10 years. Quantification of vascular risk is accomplished through risk equations or risk score sheets that have been developed on the basis of observations from large cohort studies. For example, the Framingham risk score has been routinely applied, validated and calibrated for use. However, CVD risk prediction is an ongoing work in progress and new risk factors or markers are being identified and proposed constantly. The critical question arises is to how to evaluate the usefulness of a new marker? Four initial decisions that guide the process are:

- defining the population of interest
- defining the outcome of interest
- choosing how to incorporate the competing pre-existing set of risk factors
- selecting the appropriate model and tests to evaluate the incremental yield of a new biomarker

Since, none of the numerous new markers proposed comes close in magnitude to the necessary levels of association, some have argued that we need to wait for new and better markers; others have sought model performance measures beyond the AUC calculated from a ROC curve to evaluate the usefulness of markers. For example, the Net Reclassification Index (or Improvement-NRI), focuses on reclassification tables constructed separately for participants with and without events, and quantifies the correct movement in categories – upwards for events and downwards for non-events. In its simplest terms, the NRI is defined as a measure of the net % of those who do or do not develop an endpoint within a given time period that are correctly reclassified to a different category when a new risk factor is added to the risk estimation [1]. Again in its simplest terms, one can construct a 2×2 table and assess an endpoint, then add a new risk factor and reassess. The % improvement in TP and TN is the NRI. One example of this is the use of the coronary artery calcium (CAC) score to reclassify the patients risk say from that predicted by the FRS. The addition of a CAC score in one study, altered conventional risk determination (Framingham Risk Score [FRS]) such that the posttest probability could reclassify a patient to a new category of risk.

Although the data using the NRI are conceptually appealing for patient care, there are still many unanswered questions with substantial clinical implications that will need to be addressed prior to using this reclassification in clinical practice.

Screening Testing

Screening tests are ubiquitous in contemporary practice, yet the principles of screening are widely misunderstood. Screening is the testing of apparently well people to find those at increased risk of having a disease or disorder. Those identified are sometimes then offered a subsequent diagnostic test or procedure, or, in some instances, a treatment or preventive medication. Looking for additional illnesses in those with medical problems is termed case finding. Although an earlier diagnosis generally has intuitive appeal, earlier might not always be better, or worth the cost. For tests with continuous variables – e.g., blood glucose – sensitivity and specificity as mentioned prior, are inversely related; where the cutoff for abnormal is placed should indicate the clinical effect of wrong results. As also prior mentioned, the prevalence of disease in a population affects screening test performance: in low-prevalence settings, even very good tests have poor positive predictive values. Hence, knowledge of the approximate prevalence of the index disease is a prerequisite to interpreting screening test results.

Screening differs from the traditional clinical use of tests in several important ways. Ordinarily, patients consult with clinicians about complaints or problems; and, this prompts testing to confirm or exclude a diagnosis. Because the patient is in pain and requests help, the risk and expense of tests are usually deemed acceptable by the patient. By contrast, screening engages apparently healthy individuals who are not seeking medical help (and who might prefer to be left alone). Hence, the cost, injury, and stigmatization related to screening are especially important (though often ignored in our zeal for earlier diagnosis). Furthermore, the medical and ethical standards of screening should be, correspondingly, higher than with diagnostic tests. Bluntly put: every adverse outcome of screening is iatrogenic and entirely preventable; thus, screening has a darker side that is often overlooked.

Guidelines for Publishing or Assessing Research in Diagnostic Tests

Finally, just as there are guidelines for publishing and assessing published articles addressing clinical and observational trials (see Chaps. 3 and 19) there are also guidelines for publishing studies of new diagnostic tests. McReid et al. have suggested seven methodological standards for diagnostic tests [2] as follows.

- *Spectrum Composition*: i.e. if one changes the population under study one can change the tests diagnosticity, thus in assessing the results of a new diagnostic test, information on age and sex distribution, presenting symptoms and/or disease stage, and eligibility criteria for study patients should be included in published works.
- *Pertinent Subgroups*: Se and Sp represent average values for a population. Unless the condition is narrowly defined, the indices may vary for different medical subgroups, thus these subgroups should be clearly described.

- *Avoidance of Workup Bias*: patients with a positive or negative "gold standard" diagnostic tests might be preferentially referred to evaluate the diagnosticity of a newly reported test. For example, a new DNA test to detect the breast cancer gene was administered to biopsy proven breast cancer and cancer-free controls. Since the biopsy may be ordered preferentially in women with a family history of breast cancer, the cases selected for the new test will be enriched by a clinical factor that itself may be associated with the new DNA test.
- *Avoidance of Review Bias*: The new test needs to be interpreted independently of other tests, and the new test and the gold standard test need to be interpreted separately by persons unaware of the results of the other (akin blinding in clinical trials).
- *Precision of Results for Test Accuracy*: Like any other research, point estimates should have confidence limits reported.
- *Presentation Of Indeterminate Results*: Not all tests come out Yes or No. Sometimes they are equivocal or indeterminate. The frequency of these results may limit the tests applicability, or make it cost more because additional test are then needed. Finally,
- *Test Reproducibility*: must be reported.

References

1. Bayes T. An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond B Biol Sci. 1763;53:370–418.
2. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science. 1959;130:9–21.
3. Redwood DR, Borer JS, Epstein SE. Whither the ST segment during exercise. Circulation. 1976;54:703–6.
4. Rifkin RD, Hood Jr WB. Bayesian analysis of electrocardiographic exercise stress testing. N Engl J Med. 1977;297:681–6.
5. McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, For GG, et al. Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). CMAJ. 2004;171:1369–73. PMC527344.
6. Green DM, Swets JM. Signal detection theory and psychophysics. New York: Wiley; 1966.
7. Maisel AS, Krishnaswamy P, Nowak RM, McCord J, Hollander JE, Duc P, Omland T, Storrow AB, Abraham WT, Wu AH, Clopton P, Steg PG, Westheim A, Knudsen CW, Perez A, Kazanegra R, Herrmann HC, McCullough PA, Breathing Not Properly Multinational Study, I. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. N Engl J Med. 2002;347:161–7. doi:10.1056/NEJMoa020233.
8. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med. 1975;293:257. doi:10.1056/NEJM197507312930513.
9. Miller DD, Shaw LJ. Coronary artery disease: diagnostic and prognostic models for reducing patient risk. J Cardiovasc Nurs. 2006;21:S2–16; quiz S17–9.