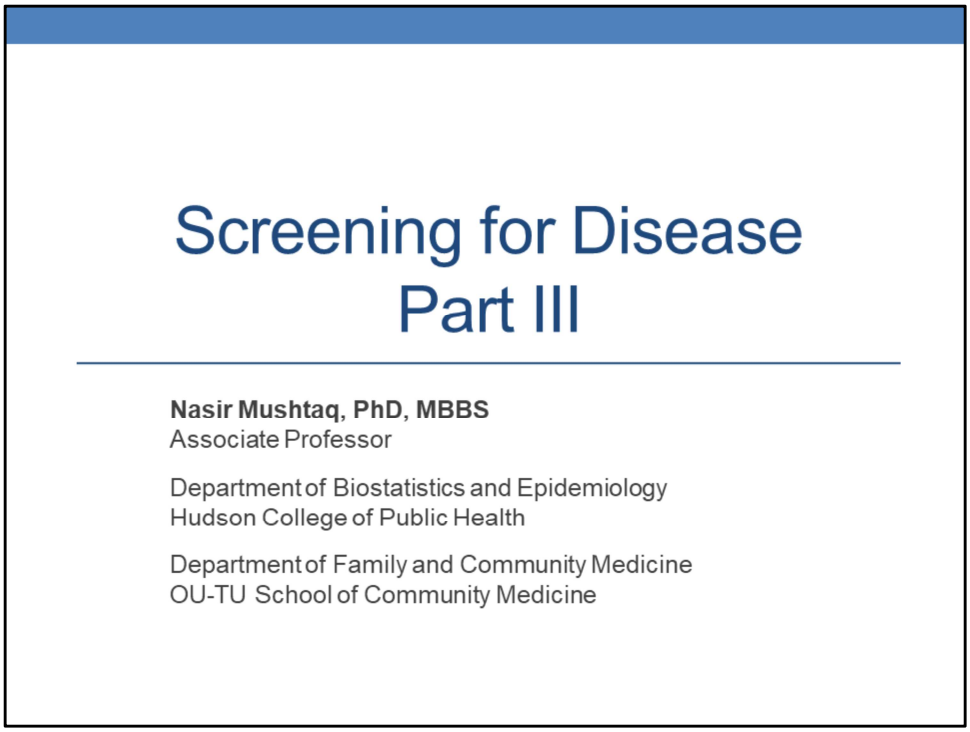
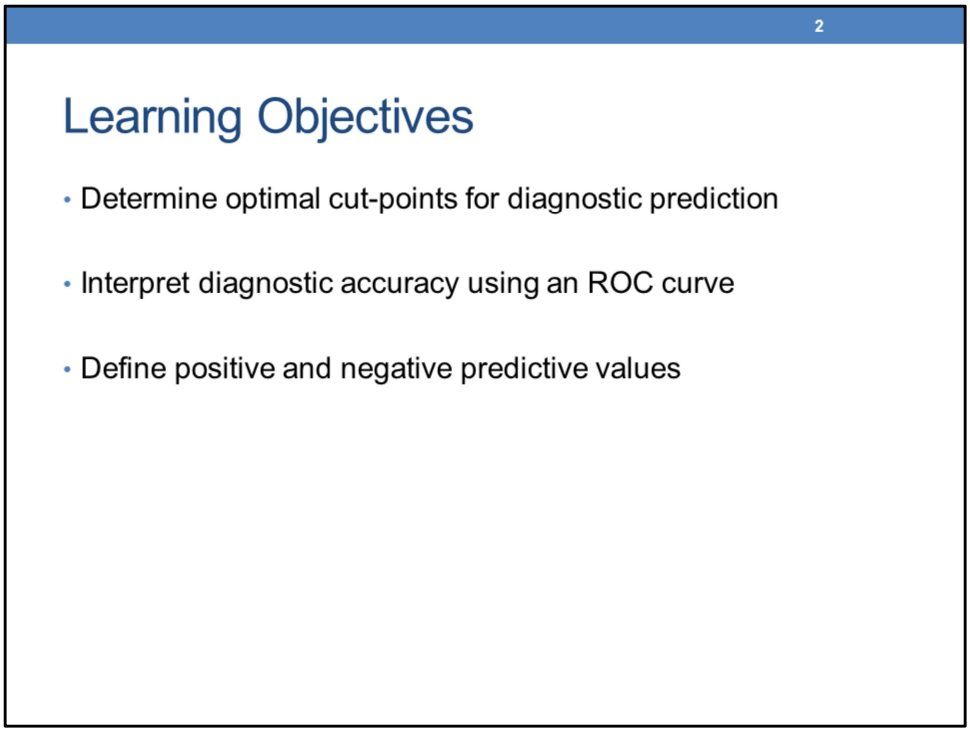
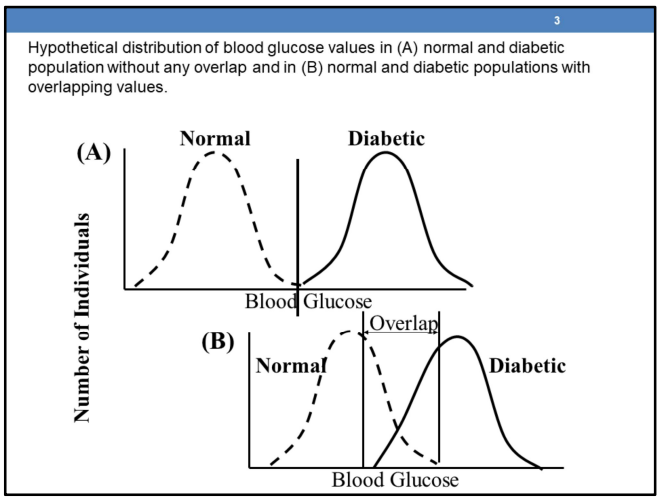


In the third part of this series focused on epidemiologic and biostatistical methods related to disease screening , we will define and learn how to compute evaluation summary measures for a screening test with a continuous measure and will also learn how to calculate and interpret the positive and negative predictive values.



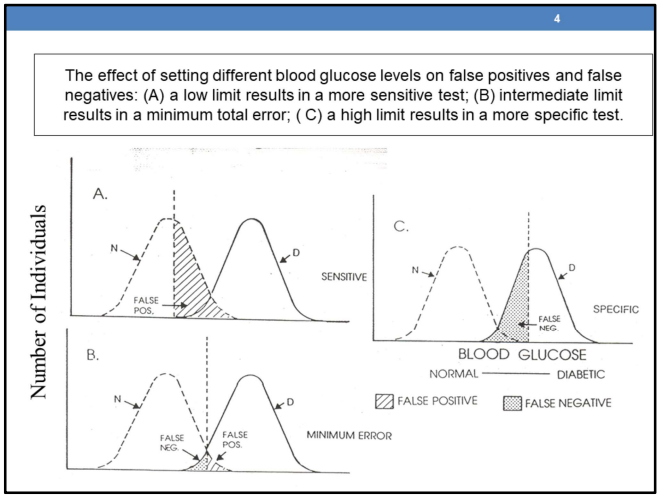
After viewing this section, you will be able to determine approaches for identifying the optimal cut-points for diagnostic prediction when using a continuous result from a screening or diagnostic test. You will be able to interpret the results from a receiver operating characteristic curve when identifying cut-points for a continuous result from a screening or diagnostic test and will be able to calculate and interpret the positive and negative predictive values from a screening or diagnostic test.



To begin our discussion, let's consider a continuous measure, blood glucose, that will be used to screen patients for diabetes. In practice, we will need to identify a cut-point for glucose that most accurately classifies participants as having diabetes or not.

In Figure A, there is no overlap in the distribution of blood glucose between those with and without diabetes. In this case, we would use the cut-point above which all participants with values above had diabetes and all participants below did not have diabetes. This cut-point would be 100% accurate.

In practice, non-overlapping distributions are not common and instead, we see overlapping distributions as in Figure B. There is a range of blood glucose values that are seen by participants with and without diabetes. In this case, the choice of the screening threshold to define positive and negative results will have an impact on accuracy. Setting a value that is too high will result in coding as negative some patients with diabetes and setting a threshold that is too low will result in coding as positive some patients who do not have diabetes.

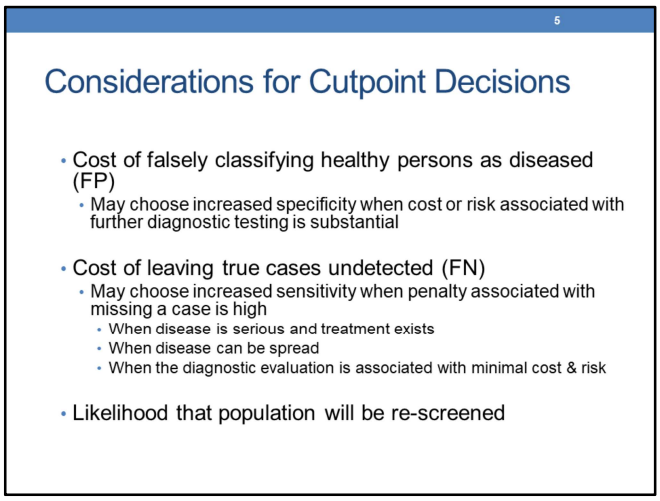


This series of figures presents a range of scenarios based on the selected cut-point to define positive and negative test results.

In Figure A, a lower threshold is used and this results in a higher false positive fraction (patients without diabetes coded as having a positive test result) and a very low false negative fraction.

In Figure C, a higher threshold is used and this results in a higher false negative fraction (patients with diabetes coded as having a negative test result) and a very low false positive fraction.

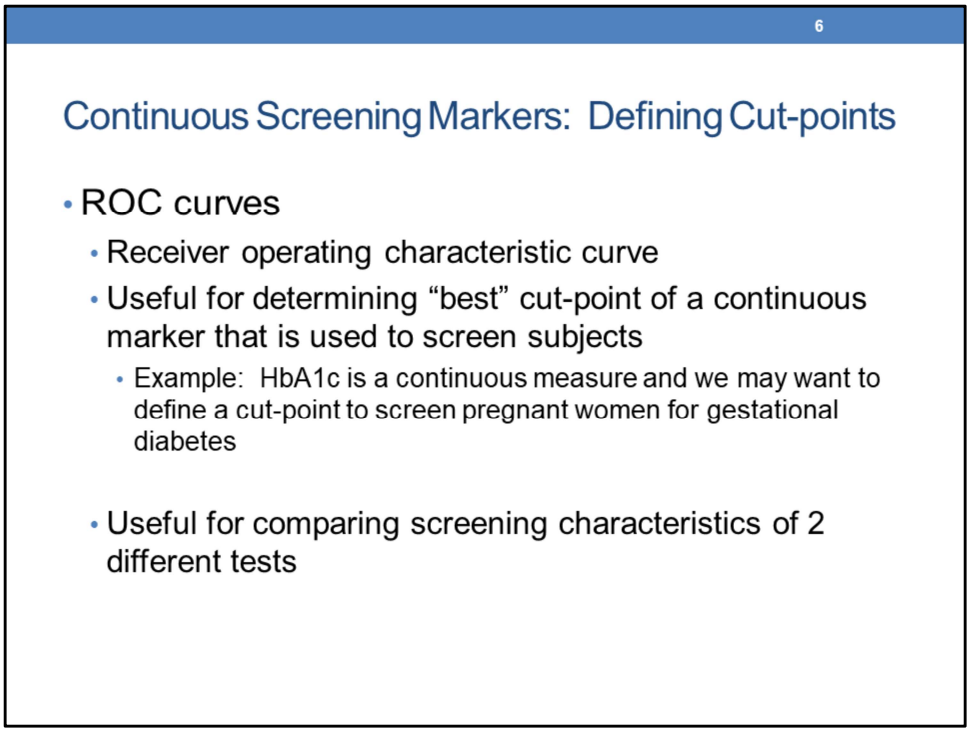
The optimal cut-point, shown in Figure B, minimizes both the false positive fraction and the false negative fraction.



When selecting a cut-point of a continuous measure to use in a screening test, we need to consider the cost of falsely classifying health persons as disease, meaning, the false positive fraction. We may want to maximize the specificity, and minimize the false positive fraction, in situations where the cost or risk associated with further diagnostic testing is substantial.

On the other hand, we may want to maximize the sensitivity, and minimize the false negative fraction, in situations where the penalty associated with missing a true case is high. For example, we would want to maximize the sensitivity in situations where the disease is serious and effective treatment exists, in situations where the disease can spread, and in situations where the diagnostic evaluation, used to follow up on positive screening results, is associated with minimal cost and risk.

In selecting the cut-point to define positive and negative cases, we will also want to consider the likelihood that the population will be rescreened for followed with more definitive diagnostic tests.



When considering cut-points for a continuous measure to define positive and negative test results, we can utilize a receive operating characteristic (ROC) curve to summarize the trade-offs between the true positive fraction (sensitivity) and the false positive fraction (1 – specificity). An ideal test has a high true positive fraction and a low false positive fraction. The ROC is a plot of the false positive fraction on the horizontal axis by the true positive fraction on the vertical axis where these values are determined for each possible cut-point of the continuous measure.

For example, we may want to identify the optimal cut-point for HbA1c to use to screen pregnant women for gestational diabetes. We would like to use a cut-point that corresponds to a high true positive fraction and a low false positive fraction.

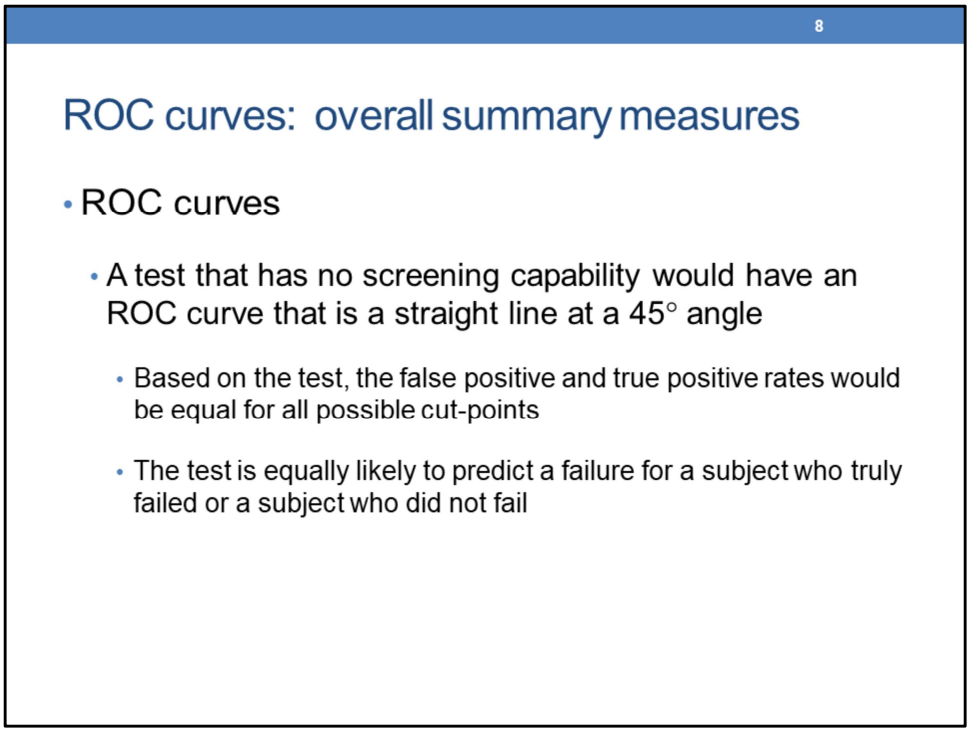
The ROC curve is also a useful methodology for comparing the performance of two screening tests.

ROC curves: statistics of interest

- ROC curves
  - For each possible cut-point, plot the sensitivity (y-axis) by 1-specificity (x-axis) [could be interpreted as a plot of the true positive by false positive rate]
  - If costs of a false positive and false negative are equal, the best cut-point will correspond to the upper, left-most point of the curve

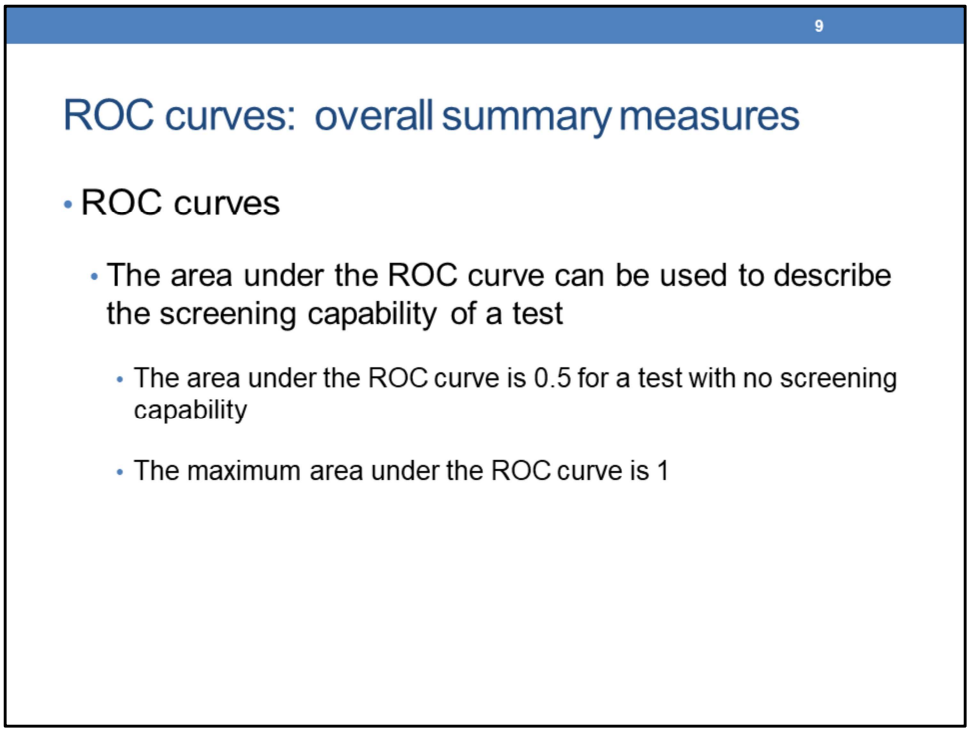
An ROC curve is created by calculating the sensitivity and specificity for each possible cut-point of the continuous measure. We then create a curve by joining the resulting points.

If the costs associated with a false positive and a false negative result are equal, we will choose the cut-point that corresponds to the upper, left-most point on the curve.

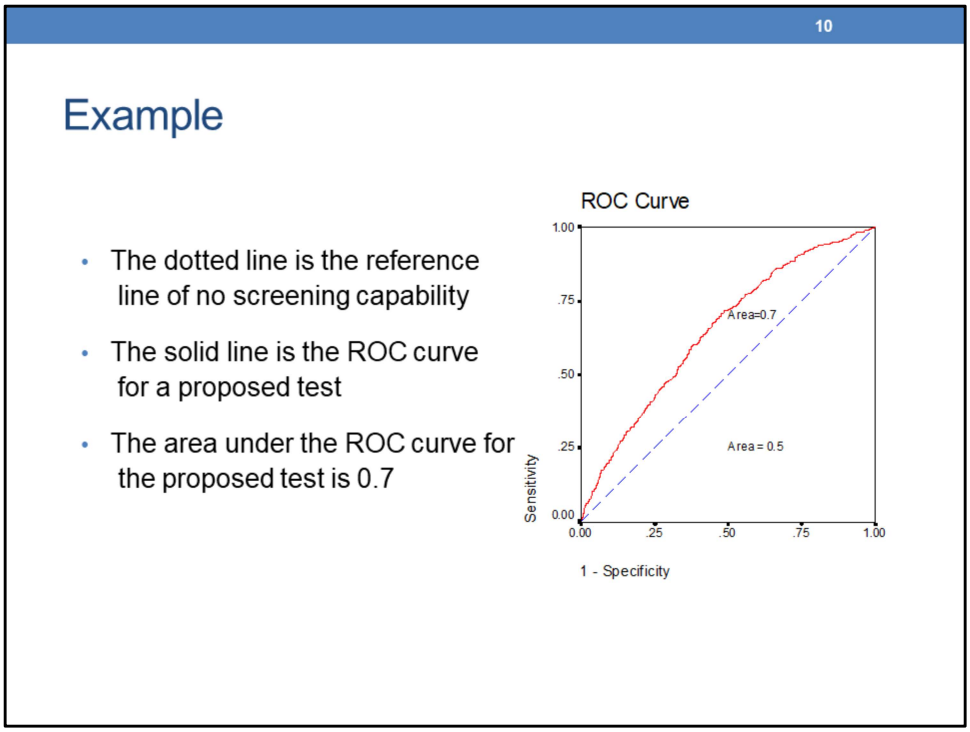


When summarizing the overall screening capability of the measure, a test with no screening utility is one that has an ROC curve which is a straight 45 degree line. This means that for each cut-point, the probability of a true positive is equal to the probability of a false positive result. Or, in other words, we would be equally likely to predict a failure for a participant who truly failed or a participant who did not fail. This would be analogous to flipping a coin to define a participant as positive or negative.

Curves that are located towards the upper left side of the grid, corresponding to higher true positive and lower false positive fraction values, are preferred in practice.

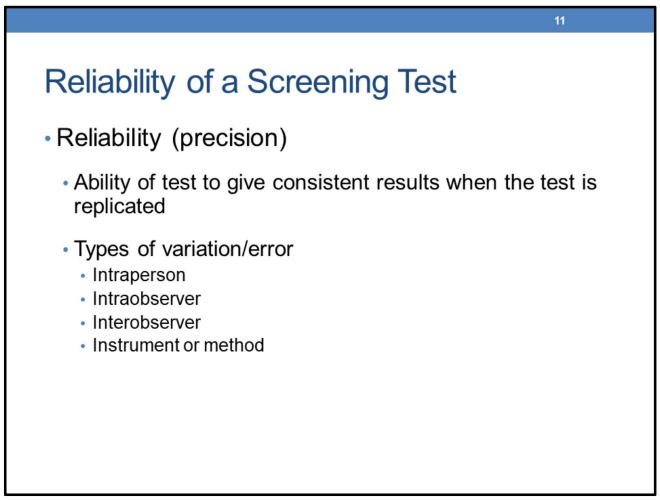


The area under the ROC curve is a useful overall summary measure of the screening capability of the test. An area of 0.5 would be seen for a test with no screening capability (i.e., corresponding to a test with an ROC curve that is equal to the 45 degree diagonal line). In contrast, a curve that has good screening capability will have an area under the ROC curve closer to a value of 1.



This figure includes an example ROC curve. The y-axis corresponds to the true positive fraction (sensitivity) and the x-axis corresponds to the false positive fraction (1-specificity). The diagonal, 45 degree reference line, corresponds to a test of no screening utility, i.e., having a performance no better than flipping a coin. A useful test will have an ROC curve that lies to the upper left of the diagonal reference line.

In this case, the red ROC curve, corresponding to the test's performance, lies above the diagonal reference line. The area under the ROC curve is 0.7, which is higher than the useless test with an area under the curve of 0.5.

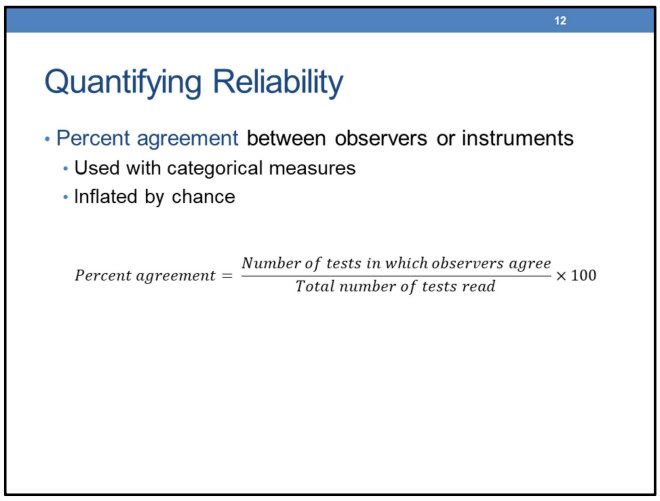


Up to this point in this series, we have focused on measures of accuracy (i.e., correctly distinguishing between those with and without disease).

Now, let's consider measures of the reliability of a test. Reliability of the test relates to the ability of the test to give consistent results with the test is replicated.

We can consider different sources of variation or error in the replication of test results. For example, we will consider the following sources:

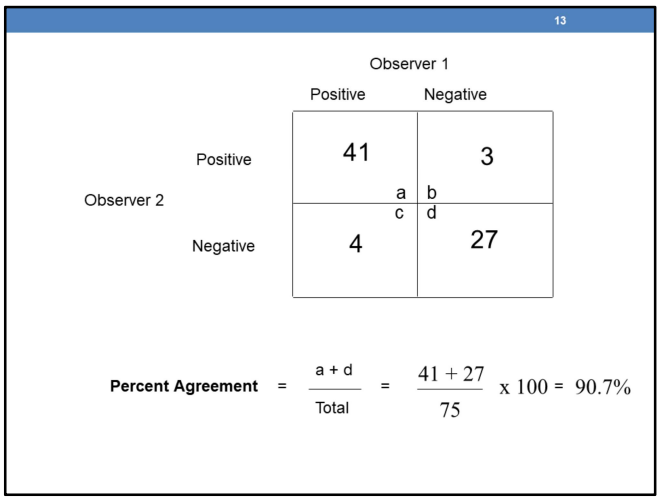
- Within-person (intra-person) variability where the person being evaluated does not remain constant in their disease status. For example, blood pressure may fluctuate throughout the day and differ between the right and left arms of a given patient.
- Within-observer (intra-observer) variability where different results are found when a given evaluator performs the test again. For example, a given evaluator may classify an X-ray as positive at one setting and as negative on another setting.
- Between-observer (inter-observer) variability where different results are found depending on who evaluates the results of the test. For example, reviewer A may identify an image as positive while reviewer B may classify the image as negative.
- Instrument or method variability where the results differ due to variation in the instrument or method itself. For example, there may be batch-to-batch variability in an assay.



A simple measure of reliability is the percent agreement between observers or instruments. The percent agreement is calculated as the number of tests in which observers agree divided by the total number of tests read. This value is then multiplied by 100%. This measure can be calculated for tests that are based on a categorical classification (e.g., classifying an image as positive or negative).

A drawback of this method is that the measure of agreement is inflated by chance agreement.

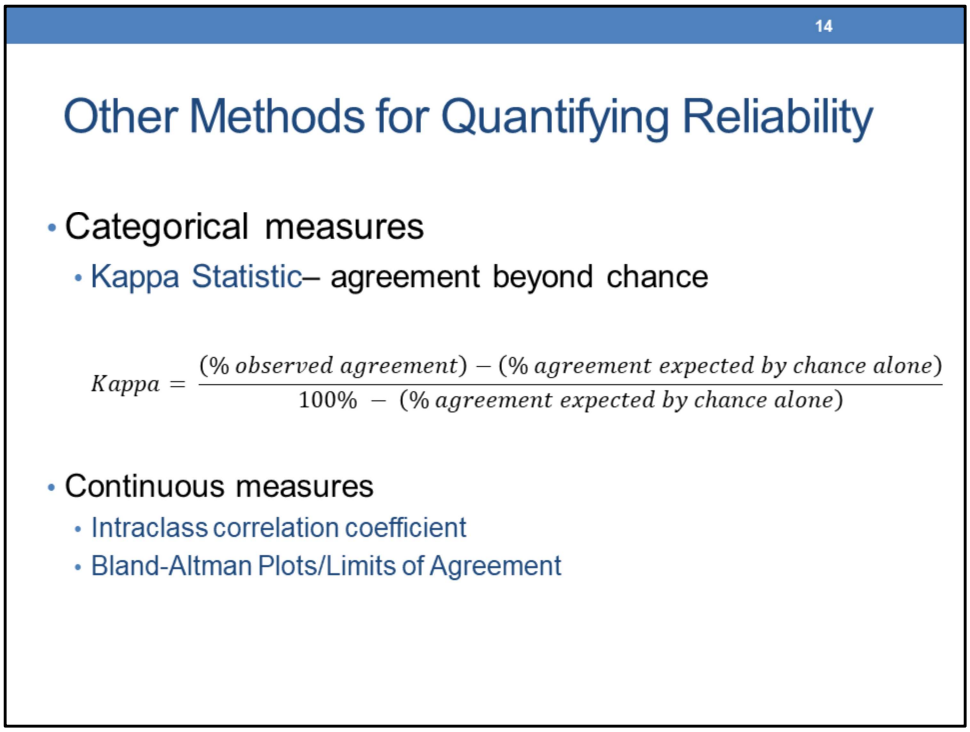
Instead, we would like to calculate a measure of agreement that is above and beyond the agreement that we would expect to see by chance alone.



Let's consider a data example in which two observers evaluate 75 images and code them as positive or negative.

The images with perfect agreement are the 41 images that were coded as positive by both observers and the 27 images that were coded as negative by both observers.

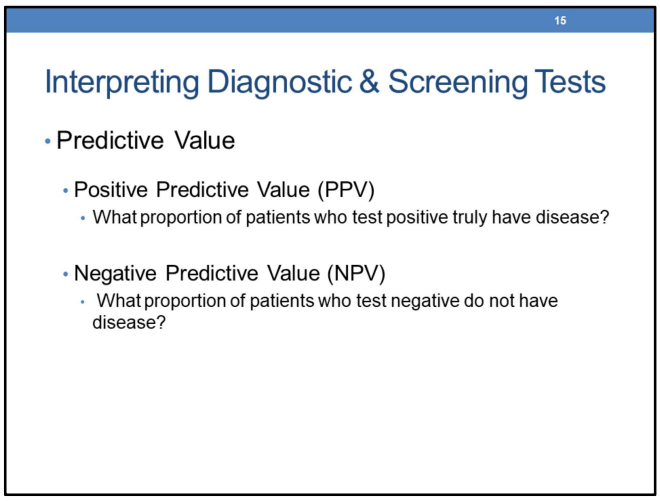
The percent agreement is therefore (68/75)*100 = 90.7%.



Other methods for quantifying reliability include the calculation of the Kappa statistic for categorical measures where the observed agreement is decreased by the percent agreement that is expected purely by chance.

If the measure is continuous instead of categorical, an intraclass correlation coefficient can be calculated to quantify the amount of variability in the measures that can be explained by between-participant variability, which as this increases towards 1 there is less variability due to other sources of variability including between-reader, within-reader, or within-participant sources of variation.

A Bland-Altman plot, which is a plot of the difference between repeated measures on the vertical axis by the average of the repeated measures on the horizontal access, can be created to summarize the limits of agreement between continuous test and re-test measures.

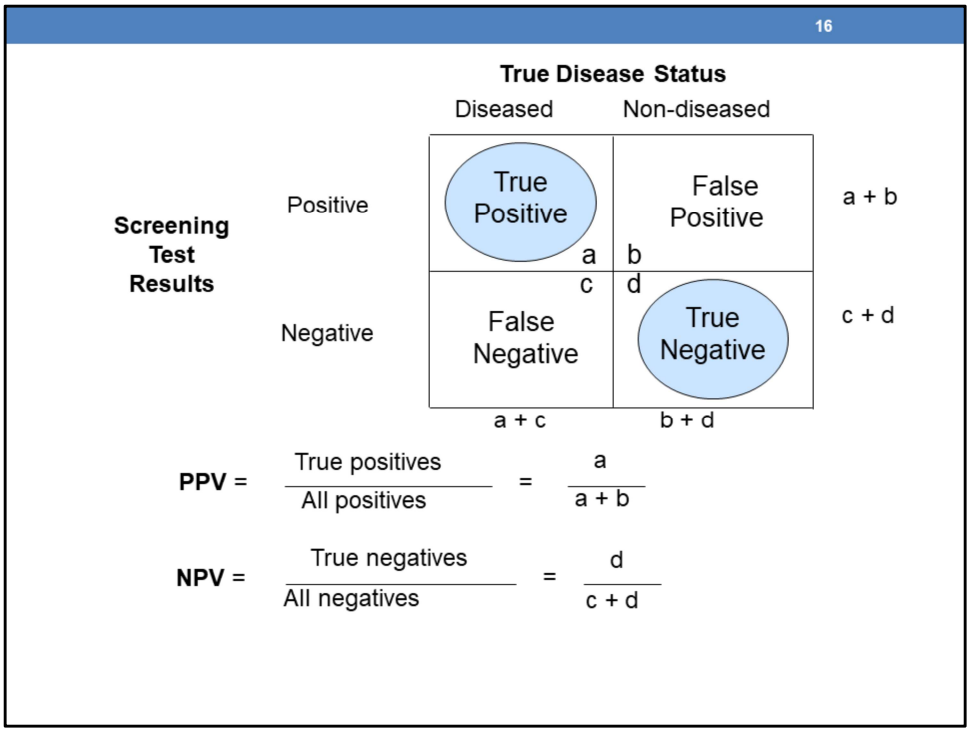


Now, let's shift our focus to the interpretation of diagnostic and screening test results.

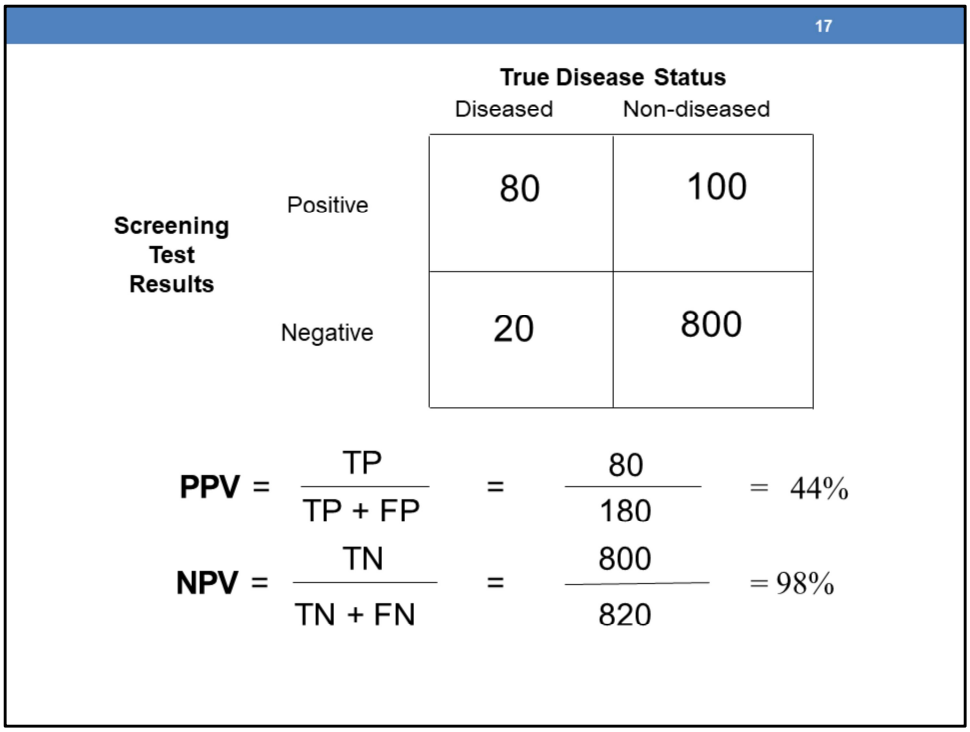
As a patient, we have a test result and based on that result, we would like to know how likely is it that I have the disease or do not have the disease. Predictive values can be used to address these questions.

The positive predictive value is the probability of having the disease given that you received a positive test result. This is conditioned or calculated among the subgroup who has a positive test result. The positive predictive value is the proportion who truly have the disease among those who tested positive.

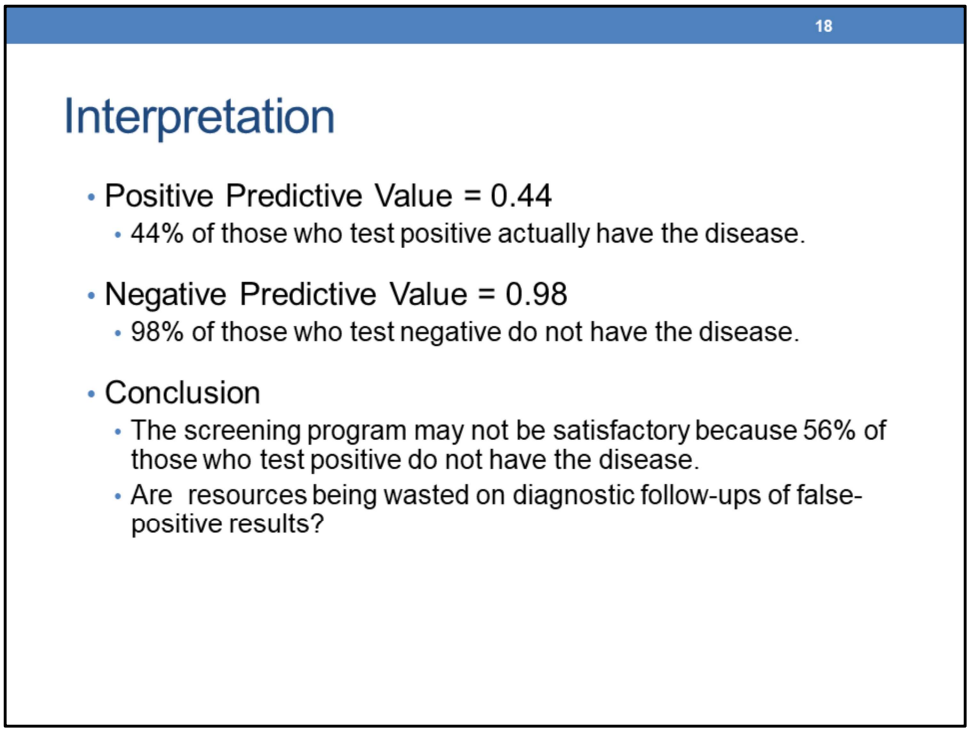
The negative predictive value is the probability of not having the disease given that the patient received a negative test result. This is conditioned or calculated among the subgroup who has a negative test result. The negative predictive value is the proportion who do not have the disease among those who tested negative.



If we return to the notation from our 2 by 2 table relating true disease status to the results of the screening test, we see that the formula for the positive predictive value is the number with disease among those who tested positive, or (a) divided by (a+b). The negative predictive value is the number without disease among those who tested negative or (d) divided by (c+d). We see that for these calculations, the denominators reflect the outcome of the tests, either positive or negative, whereas the denominators for sensitivity and specificity reflected the true disease status, either with or without disease.



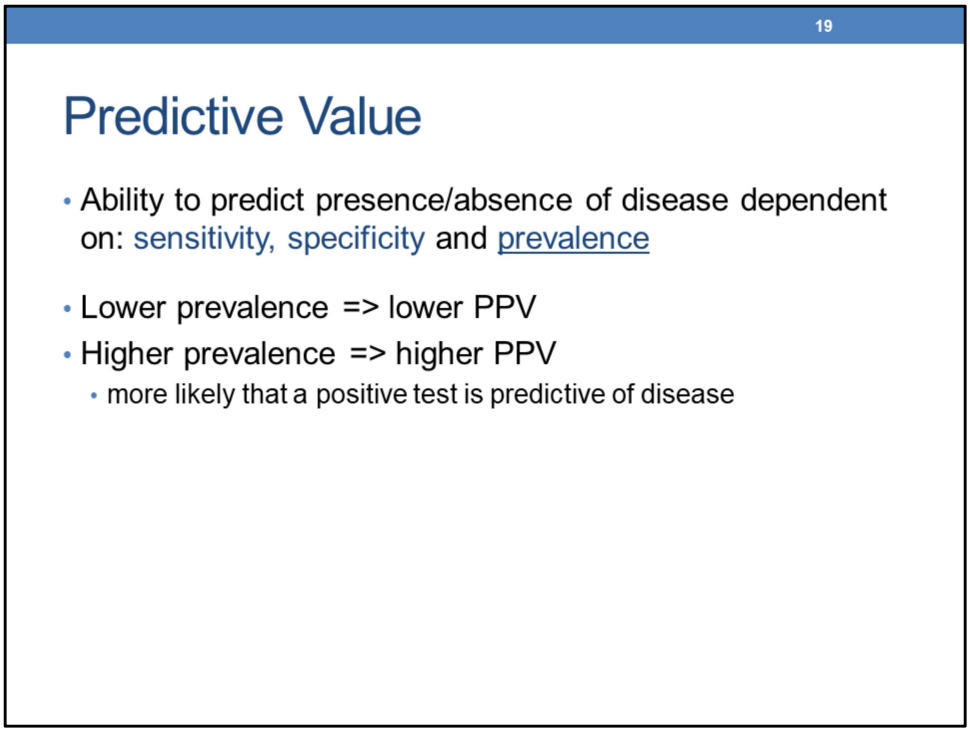
If we return to our data example from the previous segment of this series, we see that the positive predictive value is 80 diseased among 180 test positive results, corresponding to a positive predictive value of 44%. Similarly, the negative predictive value is found by taking the 800 non-diseased among the 820 negative results, corresponding to a negative predictive value of 98%.



With a positive predictive value of 44%, we can say that only 44% of those with a positive test result actually have the disease.

With a negative predictive value of 98%, we can say that 98% of those with a negative test result do not have the disease.

In conclusion, the screening program may not be satisfactory because 56% of those who test positive do not have the disease. We are concerned that with this test we may be wasting resources on diagnostic follow-up tests among those with a false positive result.



It is important to keep in mind that the positive and negative predictive values are dependent on the prevalence of disease. If you apply the screening test in a population with a lower prevalence of disease, the positive predictive value will be lower. If you apply the screening test in a population with a higher prevalence of disease, the positive predictive value will be higher.

Demonstration of How Prevalence Effects PPV

Screening test with 90% sensitivity and 90% specificity in population of 1,000.

If prevalence = 1%

		Truth		Total
		+	-	
Test	+	9	99	108
	-	1	891	892
Total		10	990	1000

$$PPV = \frac{9}{108} = 0.08$$

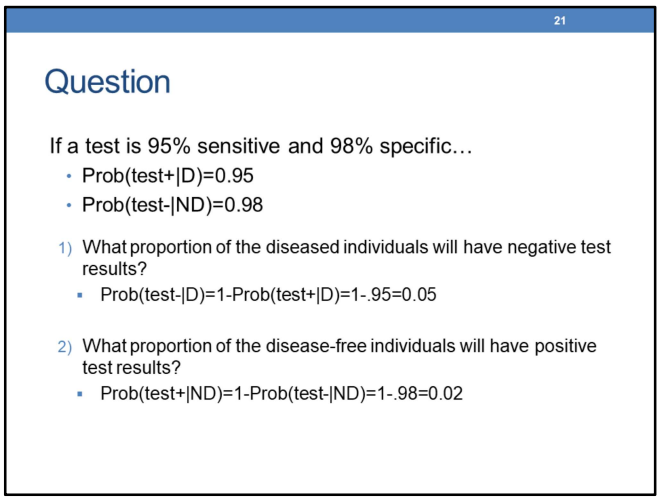
If prevalence = 5%

		Truth		Total
		+	-	
Test	+	45	95	140
	-	5	855	860
Total		50	950	1000

$$PPV = \frac{45}{140} = 0.32$$

To see the dependence of the positive predictive value on the prevalence of disease, consider these two data scenarios. In each case, the screening test is 90% sensitive and the test is 90% specific. In the first scenario, the prevalence of disease is 1% while in the second scenario, the prevalence of disease is 5%.

Even though the sensitivity and specificity of the test do not change, we see that the positive predictive value is higher in the scenario with the higher prevalence of disease. It is more likely in the setting with a higher prevalence that a patient with a positive test will actually have the disease because the disease is more common.

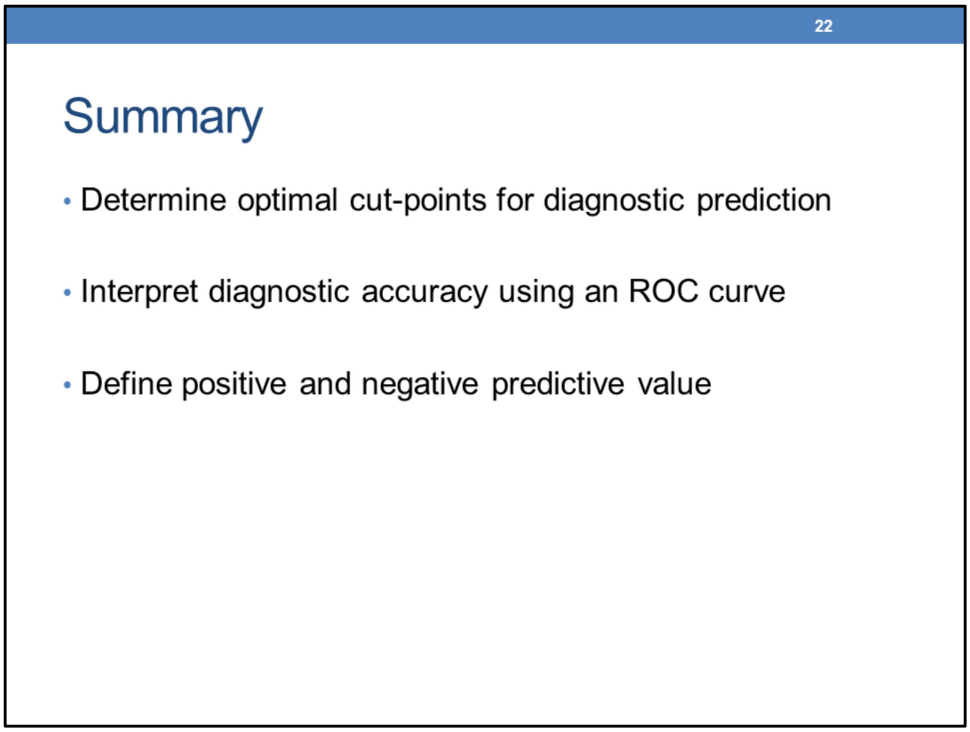


Now, let's consider a review question.

Consider a test that is 95% sensitive and 98% specific.

What proportion of the diseased individuals will have a negative test result?
Answer: we know that 95% of the diseased individuals will have a positive test result based on the value of sensitivity; therefore, 5% of the diseased individuals will have a negative test result.

What proportion of the disease-free individuals will have a positive test result?
Answer: we know that 98% of those without disease will have a negative test result based on the value of specificity; therefore, 2% of the non-diseased participants will have a positive test result.



In summary, we have reviewed considerations for defining an optimal cut-point for a continuous measure that can be used to classify participants as positive or negative for a screening test and corresponding trade-offs between false positive and true positive results.

We introduced an ROC curve as a useful summary tool for displaying the trade-offs between the true positive fraction and the false positive fraction when considering a continuous screening measure.

Finally, we defined and learned how to calculate and interpret the positive and negative predictive values, which provide estimated probabilities of disease among those who screen positive or negative.

In the final segment of this series, we will discuss sources of bias that may impact our evaluation of screening tests.