Wikipedia
The Free Encyclopedia

Sensitivity and specificity

In medicine and statistics, *sensitivity* and *specificity* mathematically describe the accuracy of a test that reports the presence or absence of a medical condition. If individuals who have the condition are considered "positive" and those who do not are considered "negative", then sensitivity is a measure of how well a test can identify true positives and specificity is a measure of how well a test can identify true negatives:

- **Sensitivity** (true positive rate) is the probability of a positive test result, conditioned on the individual truly being positive.
- **Specificity** (true negative rate) is the probability of a negative test result, conditioned on the individual truly being negative.

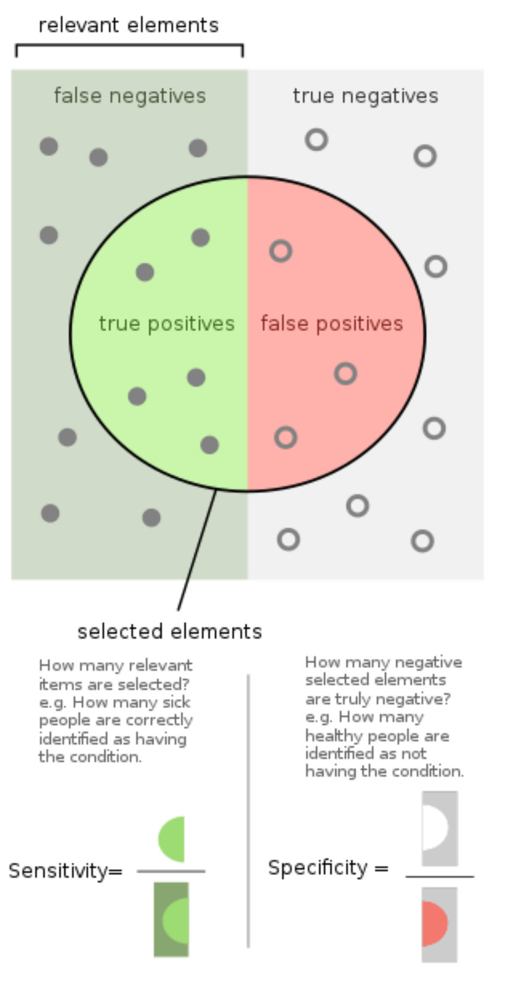
If the true status of the condition cannot be known, sensitivity and specificity can be defined relative to a "gold standard test" which is assumed correct. For all testing, both diagnoses and screening, there is usually a trade-off between sensitivity and specificity, such that higher sensitivities will mean lower specificities and vice versa.

A test which reliably detects the presence of a condition, resulting in a high number of true positives and low number of false negatives, will have a high sensitivity. This is especially important when the consequence of failing to treat the condition is serious and/or the treatment is very effective and has minimal side effects.

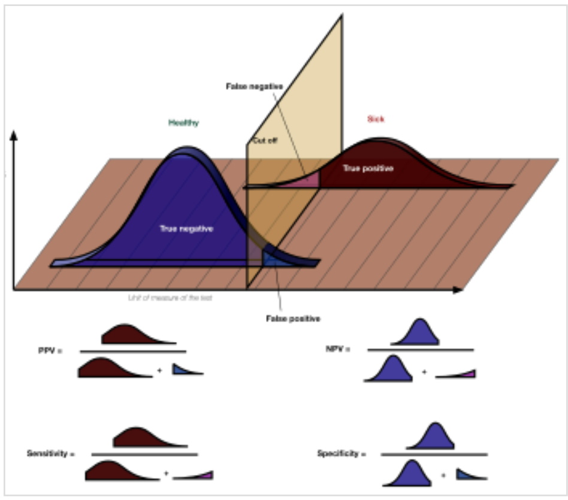
A test which reliably excludes individuals who do not have the condition, resulting in a high number of true negatives and low number of false positives, will have a high specificity. This is especially important when people who are identified as having a condition may be subjected to more testing, expense, stigma, anxiety, etc.

The terms "sensitivity" and "specificity" were introduced by American biostatistician Jacob Yerushalmy in 1947.[1]

There are different definitions within laboratory quality control, wherein "analytical sensitivity" is defined as the smallest amount of substance in a sample that can accurately be measured by an assay (synonymously to detection limit), and "analytical specificity" is defined as the ability of an assay to measure one particular organism or substance, rather than others.[12] However, this article deals with diagnostic sensitivity and specificity as defined at top.


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.
How many negative selected elements are truly negative? e.g. How many healthy people are identified as not having the condition.
Sensitivity=
Specificity =

Sensitivity and specificity - The left half of the image with the solid dots represents individuals who have the condition, while the right half of the image with the hollow dots represents individuals who do not have the condition. The circle represents all individuals who tested positive.



Sensitivity and specificity

Application to screening study

Imagine a study evaluating a test that screens people for a disease. Each person taking the test either has or does not have the disease. The test outcome can be positive (classifying the person as having the disease) or negative (classifying the person as not having the disease). The test results for each subject may or may not match the subject's actual status. In that setting:

- True positive: Sick people correctly identified as sick
- False positive: Healthy people incorrectly identified as sick
- True negative: Healthy people correctly identified as healthy
- False negative: Sick people incorrectly identified as healthy

After getting the numbers of true positives, false positives, true negatives, and false negatives, the sensitivity and specificity for the test can be calculated. If it turns out that the sensitivity is high then any person who has the disease is likely to be classified as positive by the test. On the other hand, if the specificity is high, any person who does not have the disease is likely to be classified as negative by the test. An NIH web site has a discussion of how these ratios are calculated.[13]

Definition

Sensitivity

Consider the example of a medical test for diagnosing a condition. Sensitivity (sometimes also named the detection rate in a clinical setting) refers to the test's ability to correctly detect ill patients out of those who do have the condition.[14] Mathematically, this can be expressed as:

$$\begin{aligned}\text{sensitivity} &= \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}\\ &= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}}\\ &= \text{probability of a positive test given that the patient has the disease}\end{aligned}$$

A negative result in a test with high sensitivity can be useful for "ruling out" disease,[14] since it rarely misdiagnoses those who do have the disease. A test with 100% sensitivity will recognize all patients with the disease by testing positive. In this case, a negative test result would definitively *rule out* the presence of the disease in a patient. However, a positive result in a test with high sensitivity is not necessarily useful for "ruling in" disease. Suppose a 'bogus' test kit is designed to always give a positive reading. When used on diseased patients, all patients test positive, giving the test 100% sensitivity. However, sensitivity does not take into account false positives. The bogus test also returns positive on all healthy patients, giving it a false positive rate of 100%, rendering it useless for detecting or "ruling in" the disease.

The calculation of sensitivity does not take into account indeterminate test results. If a test cannot be repeated, indeterminate samples either should be excluded from the analysis (the number of exclusions should be stated when quoting sensitivity) or can be treated as false negatives (which gives the worst-case value for sensitivity and may therefore underestimate it).

A test with a higher sensitivity has a lower type II error rate.

Specificity

Consider the example of a medical test for diagnosing a disease. Specificity refers to the test's ability to correctly reject healthy patients without a condition. Mathematically, this can be written as:

$$\begin{aligned}\text{specificity} &= \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}\\ &= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}}\\ &= \text{probability of a negative test given that the patient is well}\end{aligned}$$

A positive result in a test with high specificity can be useful for "ruling in" disease, since the test rarely gives positive results in healthy patients.[15] A test with 100% specificity will recognize all patients without the disease by testing negative, so a positive test result would definitively *rule in* the presence of the disease. However, a negative result from a test with high specificity is not necessarily useful for "ruling out" disease. For example, a test that always returns a negative test result will have a specificity of 100% because specificity does not consider false negatives. A test like that would return negative for patients with the disease, making it useless for "ruling out" the disease.

A test with a higher specificity has a lower type I error rate.

Terminology and derivations
from a confusion matrix

condition positive (P)
the number of real positive cases in the data
condition negative (N)
the number of real negative cases in the data

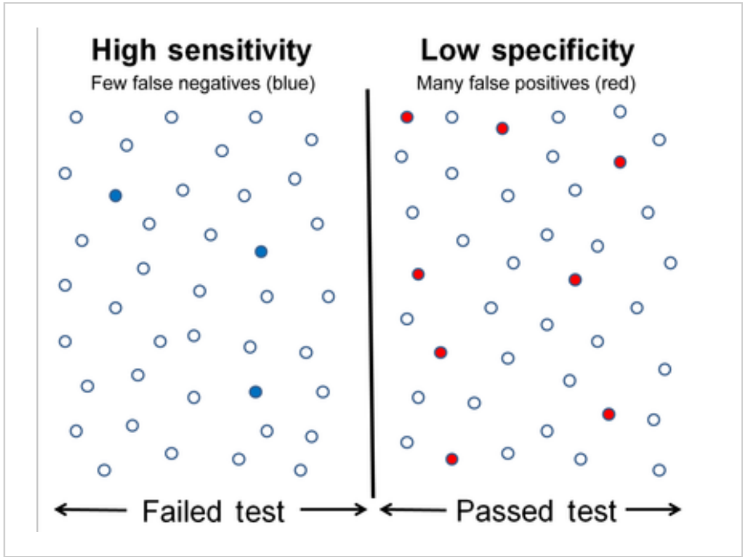
true positive (TP)
A test result that correctly indicates the presence of a condition or characteristic
true negative (TN)
A test result that correctly indicates the absence of a condition or characteristic
false positive (FP), Type I error
A test result which wrongly indicates that a particular condition or attribute is present
false negative (FN), Type II error
A test result which wrongly indicates that a particular condition or attribute is absent

sensitivity, recall, hit rate, or true positive rate (TPR)
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$
specificity, selectivity or true negative rate (TNR)
$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$
precision or positive predictive value (PPV)
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{FDR}$$
negative predictive value (NPV)
$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} = 1 - \mathrm{FOR}$$
miss rate or false negative rate (FNR)
$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} = 1 - \mathrm{TPR}$$
fall-out or false positive rate (FPR)
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$
false discovery rate (FDR)
$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}} = 1 - \mathrm{PPV}$$
false omission rate (FOR)
$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TN}} = 1 - \mathrm{NPV}$$
Positive likelihood ratio (LR+)
$$\mathrm{LR+} = \frac{\mathrm{TPR}}{\mathrm{FPR}}$$
Negative likelihood ratio (LR-)
$$\mathrm{LR-} = \frac{\mathrm{FNR}}{\mathrm{TNR}}$$
prevalence threshold (PT)
$$\mathrm{PT} = \frac{\sqrt{\mathrm{FPR}}}{\sqrt{\mathrm{TPR}} + \sqrt{\mathrm{FPR}}}$$
threat score (TS) or critical success index (CSI)
$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}$$

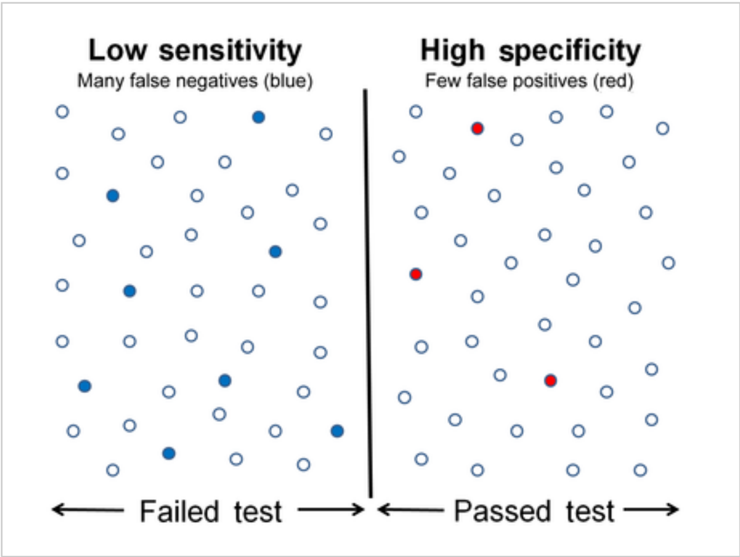
Prevalence
$$\frac{\mathrm{P}}{\mathrm{P}+\mathrm{N}}$$
accuracy (ACC)
$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$
balanced accuracy (BA)
$$\mathrm{BA} = \frac{\mathrm{TPR}+\mathrm{TNR}}{2}$$
F1 score
is the harmonic mean of precision and sensitivity: $\mathrm{F}_1 = 2 \times \frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$
phi coefficient (φ or r_φ) or Matthews correlation coefficient (MCC)
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$
Fowlkes–Mallows index (FM)
$$\mathrm{FM} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}} = \sqrt{\mathrm{PPV} \times \mathrm{TPR}}$$
informedness or bookmaker informedness (BM)
$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1$$
markedness (MK) or deltaP (Δp)
$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1$$
Diagnostic odds ratio (DOR)
$$\mathrm{DOR} = \frac{\mathrm{LR+}}{\mathrm{LR-}}$$

Sources: Fawcett (2006),[2] Piryonesi and El-Diraby (2020),[3] Powers (2011),[4] Ting (2011),[5] CAWCR,[6] D. Chicco & G. Jurman (2020, 2021, 2023),[7][8][9] Tharwat (2018).[10] Balayla (2020)[11]

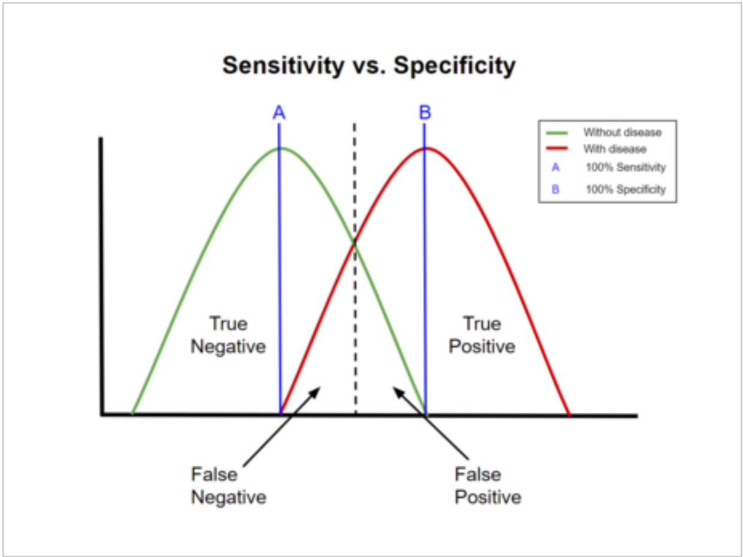
Graphical illustration



High sensitivity and low specificity



Low sensitivity and high specificity



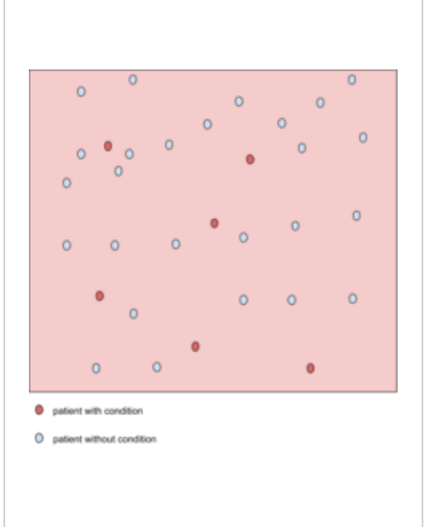
A graphical illustration of sensitivity and specificity

The above graphical illustration is meant to show the relationship between sensitivity and specificity. The black, dotted line in the center of the graph is where the sensitivity and specificity are the same. As one moves to the left of the black dotted line, the sensitivity increases, reaching its maximum value of 100% at line A, and the specificity decreases. The sensitivity at line A is 100% because at that point there are zero false negatives, meaning that all the negative test results are true negatives. When moving to the right, the opposite applies, the specificity increases until it reaches the B line and becomes 100% and the sensitivity decreases. The specificity at line B is 100% because the number of false positives is zero at that line, meaning all the positive test results are true positives.

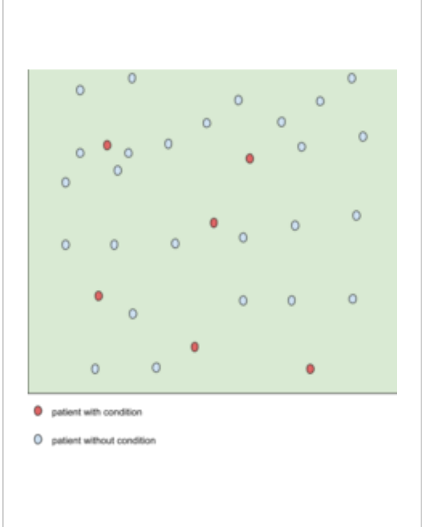
The middle solid line in both figures that show the level of sensitivity and specificity is the test cutoff point. As previously described, moving this line results in a trade-off between the level of sensitivity and specificity. The left-hand side of this line contains the data points that tests below the cut off point and are considered negative (the blue dots indicate the False Negatives (FN), the white dots True Negatives (TN)). The right-hand side of the line shows the data points that tests above the cut off point and are considered positive (red dots indicate False Positives (FP)). Each side contains 40 data points.

For the figure that shows high sensitivity and low specificity, there are 3 FN and 8 FP. Using the fact that positive results = true positives (TP) + FP, we get TP = positive results - FP, or TP = 40 - 8 = 32. The number of sick people in the data set is equal to TP + FN, or 32 + 3 = 35. The sensitivity is therefore 32 / 35 = 91.4%. Using the same method, we get TN = 40 - 3 = 37, and the number of healthy people 37 + 8 = 45, which results in a specificity of 37 / 45 = 82.2 %.

For the figure that shows low sensitivity and high specificity, there are 8 FN and 3 FP. Using the same method as the previous figure, we get TP = 40 - 3 = 37. The number of sick people is 37 + 8 = 45, which gives a sensitivity of 37 / 45 = 82.2 %. There are 40 - 8 = 32 TN. The specificity therefore comes out to 32 / 35 = 91.4%.



A test result with 100 percent sensitivity



A test result with 100 percent specificity

The red dot indicates the patient with the medical condition. The red background indicates the area where the test predicts the data point to be positive. The true positive in this figure is 6, and false negatives of 0 (because all positive condition is correctly predicted as positive). Therefore, the sensitivity is 100% (from 6 / (6 + 0)). This situation is also illustrated in the previous figure where the dotted line is at position A (the left-hand side is predicted as negative by the model, the right-hand side is predicted as positive by the model). When the dotted line, test cut-off line, is at position A, the test correctly predicts all the population of the true positive class, but it will fail to correctly identify the data point from the true negative class.

Similar to the previously explained figure, the red dot indicates the patient with the medical condition. However, in this case, the green background indicates that the test predicts that all patients are free of the medical condition. The number of data point that is true negative is then 26, and the number of false positives is 0. This result in 100% specificity (from 26 / (26 + 0)). Therefore, sensitivity or specificity alone cannot be used to measure the performance of the test.

Medical usage

In medical diagnosis, test sensitivity is the ability of a test to correctly identify those with the disease (true positive rate), whereas test specificity is the ability of the test to correctly identify those without the disease (true negative rate). If 100 patients known to have a disease were tested, and 43 test positive, then the test has 43% sensitivity. If 100 with no disease are tested and 96 return a completely negative result, then the test has 96% specificity. Sensitivity and specificity are prevalence-independent test characteristics, as their values are intrinsic to the test and do not depend on the disease prevalence in the population of interest.[16] Positive and negative predictive values, but not sensitivity or specificity, are values influenced by the prevalence of disease in the population that is being tested. These concepts are illustrated graphically in this applet Bayesian clinical diagnostic model (https://kennis-research.shinyapps.io/Bayes-App/) which show the positive and negative predictive values as a function of the prevalence, sensitivity and specificity.

Misconceptions

It is often claimed that a highly specific test is effective at ruling in a disease when positive, while a highly sensitive test is deemed effective at ruling out a disease when negative.[17][18] This has led to the widely used mnemonics SPPIN and SNNOUT, according to which a highly **sp**ecific test, when **p**ositive, rules **in** disease (SP-P-IN), and a highly **sen**sitive test, when **n**egative, rules **out** disease (SN-N-OUT). Both rules of thumb are, however, inferentially misleading, as the diagnostic power of any test is determined by the prevalence of the condition being tested, the test's sensitivity *and* its specificity.[19][20][21] The SNNOUT mnemonic has some validity when the prevalence of the condition in question is extremely low in the tested sample.

The tradeoff between specificity and sensitivity is explored in ROC analysis as a trade off between TPR and FPR (that is, recall and fallout).[22] Giving them equal weight optimizes informedness = specificity + sensitivity − 1 = TPR − FPR, the magnitude of which gives the probability of an informed decision between the two classes (> 0 represents appropriate use of information, 0 represents chance-level performance, < 0 represents perverse use of information).[23]

Sensitivity index

The sensitivity index or d' (pronounced "dee-prime") is a statistic used in signal detection theory. It provides the separation between the means of the signal and the noise distributions, compared against the standard deviation of the noise distribution. For normally distributed signal and noise with mean and standard deviations μ_S and σ_S, and μ_N and σ_N, respectively, d' is defined as:

$$d' = \frac{\mu_S - \mu_N}{\sqrt{\frac{1}{2}\left(\sigma_S^2 + \sigma_N^2\right)}}$$[24]

An estimate of d' can be also found from measurements of the hit rate and false-alarm rate. It is calculated as:

$d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$,[25]

where function $Z(p)$, $p \in [0, 1]$, is the inverse of the cumulative Gaussian distribution.

d' is a dimensionless statistic. A higher d' indicates that the signal can be more readily detected.

Confusion matrix

The relationship between sensitivity, specificity, and similar terms can be understood using the following table. Consider a group with **P** positive instances and **N** negative instances of some condition. The four outcomes can be formulated in a 2×2 *contingency table* or *confusion matrix*, as well as derivations of several metrics using the four outcomes, as follows:

Sources: [26][27][28][29][30][31][32][33][34]

		Predicted condition			
	Total population $= \mathrm{P} + \mathrm{N}$	**Predicted Positive (PP)**	**Predicted Negative (PN)**	Informedness, bookmaker informedness (BM) $= \mathrm{TPR} + \mathrm{TNR} - 1$	Prevalence threshold (PT) $= \frac{\sqrt{\mathrm{TPR} \times \mathrm{FPR}} - \mathrm{FPR}}{\mathrm{TPR} - \mathrm{FPR}}$
Actual condition	**Positive (P)**	**True positive (TP),** **hit**	**False negative (FN),** **type II error, miss,** **underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{\mathrm{TP}}{\mathrm{P}} = 1 - \mathrm{FNR}$	False negative rate (FNR), miss rate $= \frac{\mathrm{FN}}{\mathrm{P}} = 1 - \mathrm{TPR}$
	Negative (N)	**False positive (FP),** **type I error, false alarm,** **overestimation**	**True negative (TN),** **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{\mathrm{FP}}{\mathrm{N}} = 1 - \mathrm{TNR}$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{\mathrm{TN}}{\mathrm{N}} = 1 - \mathrm{FPR}$
	Prevalence $= \frac{\mathrm{P}}{\mathrm{P} + \mathrm{N}}$	Positive predictive value (PPV), precision $= \frac{\mathrm{TP}}{\mathrm{PP}} = 1 - \mathrm{FDR}$	False omission rate (FOR) $= \frac{\mathrm{FN}}{\mathrm{PN}} = 1 - \mathrm{NPV}$	Positive likelihood ratio (LR+) $= \frac{\mathrm{TPR}}{\mathrm{FPR}}$	Negative likelihood ratio (LR−) $= \frac{\mathrm{FNR}}{\mathrm{TNR}}$
	Accuracy (ACC) $= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$	False discovery rate (FDR) $= \frac{\mathrm{FP}}{\mathrm{PP}} = 1 - \mathrm{PPV}$	Negative predictive value (NPV) $= \frac{\mathrm{TN}}{\mathrm{PN}} = 1 - \mathrm{FOR}$	Markedness (MK), deltaP (Δp) $= \mathrm{PPV} + \mathrm{NPV} - 1$	Diagnostic odds ratio (DOR) $= \frac{\mathrm{LR+}}{\mathrm{LR-}}$
	Balanced accuracy (BA) $= \frac{\mathrm{TPR} + \mathrm{TNR}}{2}$	F_1 score $= \frac{2\,\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$	Fowlkes–Mallows index (FM) $= \sqrt{\mathrm{PPV} \times \mathrm{TPR}}$	Matthews correlation coefficient (MCC) $= \sqrt{\mathrm{TPR} \times \mathrm{TNR} \times \mathrm{PPV} \times \mathrm{NPV}} - \sqrt{\mathrm{FNR} \times \mathrm{FPR} \times \mathrm{FOR} \times \mathrm{FDR}}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$

A worked example

A diagnostic test with sensitivity 67% and specificity 91% is applied to 2030 people to look for a disorder with a population prevalence of 1.48%

		Fecal occult blood screen test outcome			
	Total population (pop.) = 2030	Test outcome **positive**	Test outcome **negative**	Accuracy (ACC) = (TP + TN) / pop. = (20 + 1820) / 2030 ≈ **90.64%**	F_1 score $= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ ≈ **0.174**
Patients with bowel cancer (as confirmed on endoscopy)	Actual condition positive (AP) = 30 (2030 × 1.48%)	**True positive** (TP) = 20 (2030 × 1.48% × 67%)	**False negative** (FN) = 10 (2030 × 1.48% × (100% − 67%))	True positive rate (TPR), recall, sensitivity = TP / AP = 20 / 30 ≈ **66.7%**	False negative rate (FNR), miss rate = FN / AP = 10 / 30 ≈ **33.3%**
	Actual condition negative (AN) = 2000 (2030 × (100% − 1.48%))	**False positive** (FP) = 180 (2030 × (100% − 1.48%) × (100% − 91%))	**True negative** (TN) = 1820 (2030 × (100% − 1.48%) × 91%)	False positive rate (FPR), fall-out, probability of false alarm = FP / AN = 180 / 2000 = **9.0%**	Specificity, selectivity, true negative rate (TNR) = TN / AN = 1820 / 2000 = **91%**
	Prevalence = AP / pop. = 30 / 2030 ≈ **1.48%**	Positive predictive value (PPV), precision = TP / (TP + FP) = 20 / (20 + 180) = **10%**	False omission rate (FOR) = FN / (FN + TN) = 10 / (10 + 1820) ≈ **0.55%**	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$ = (20 / 30) / (180 / 2000) ≈ **7.41**	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$ = (10 / 30) / (1820 / 2000) ≈ **0.366**
		False discovery rate (FDR) = FP / (TP + FP) = 180 / (20 + 180) = **90.0%**	Negative predictive value (NPV) = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.45%**	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$ ≈ **20.2**	

Related calculations

- False positive rate (α) = type I error = 1 − specificity = FP / (FP + TN) = 180 / (180 + 1820) = 9%
- False negative rate (β) = type II error = 1 − sensitivity = FN / (TP + FN) = 10 / (20 + 10) ≈ 33%
- Power = sensitivity = 1 − β
- Positive likelihood ratio = sensitivity / (1 − specificity) ≈ 0.67 / (1 − 0.91) ≈ 7.4
- Negative likelihood ratio = (1 − sensitivity) / specificity ≈ (1 − 0.67) / 0.91 ≈ 0.37
- Prevalence threshold = $PT = \frac{\sqrt{TPR(-TNR+1)}+TNR-1}{(TPR+TNR-1)}$ ≈ 0.2686 ≈ 26.9%

This hypothetical screening test (fecal occult blood test) correctly identified two-thirds (66.7%) of patients with colorectal cancer.[a] Unfortunately, factoring in prevalence rates reveals that this hypothetical test has a high false positive rate, and it does not reliably identify colorectal cancer in the overall population of asymptomatic people (PPV = 10%).

On the other hand, this hypothetical test demonstrates very accurate detection of cancer-free individuals (NPV ≈ 99.5%). Therefore, when used for routine colorectal cancer screening with asymptomatic adults, a negative result supplies important data for the patient and doctor, such as ruling out cancer as the cause of gastrointestinal symptoms or reassuring patients worried about developing colorectal cancer.

Estimation of errors in quoted sensitivity or specificity

Sensitivity and specificity values alone may be highly misleading. The 'worst-case' sensitivity or specificity must be calculated in order to avoid reliance on experiments with few results. For example, a particular test may easily show 100% sensitivity if tested against the gold standard four times, but a single additional test against the gold standard that gave a poor result would imply a sensitivity of only 80%. A common way to do this is to state the binomial proportion confidence interval, often calculated using a Wilson score interval.

Confidence intervals for sensitivity and specificity can be calculated, giving the range of values within which the correct value lies at a given confidence level (e.g., 95%).[37]

Terminology in information retrieval

In information retrieval, the positive predictive value is called **precision**, and sensitivity is called **recall**. Unlike the Specificity vs Sensitivity tradeoff, these measures are both independent of the number of true negatives, which is generally unknown and much larger than the actual numbers of relevant and retrieved documents. This assumption of very large numbers of true negatives versus positives is rare in other applications.[23]

The F-score can be used as a single measure of performance of the test for the positive class. The F-score is the harmonic mean of precision and recall:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In the traditional language of statistical hypothesis testing, the sensitivity of a test is called the statistical power of the test, although the word *power* in that context has a more general usage that is not applicable in the present context. A sensitive test will have fewer Type II errors.

See also

- Brier score
- Cumulative accuracy profile
- Discrimination (information)
- False positive paradox
- Hypothesis tests for accuracy
- Precision and recall
- Receiver operating characteristic
- Statistical significance
- Uncertainty coefficient, also called proficiency
- Youden's J statistic

Science portal

Biology portal

Medicine portal

Notes

a. There are advantages and disadvantages for all medical screening tests. Clinical practice guidelines, such as those for colorectal cancer screening, describe these risks and benefits.[35][36]

References

1. Yerushalmy J (1947). "Statistical problems in assessing methods of medical diagnosis with special reference to x-ray techniques". *Public Health Reports*. **62** (2): 1432–39. doi:10.2307/4586294 (https://doi.org/10.2307%2F4586294). JSTOR 4586294 (https://www.jstor.org/stable/4586294). PMID 20340527 (https://pubmed.ncbi.nlm.nih.gov/20340527). S2CID 19967899 (https://api.semanticscholar.org/CorpusID:19967899).
2. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. Bibcode:2006PaReL..27..861F (https://ui.adsabs.harvard.edu/abs/2006PaReL..27..861F). doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010). S2CID 2027090 (https://api.semanticscholar.org/CorpusID:2027090).
3. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512 (https://doi.org/10.1061%2F%28ASCE%29IS.1943-555X.0000512). S2CID 213782055 (https://api.semanticscholar.org/CorpusID:213782055).
4. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
5. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8 (https://doi.org/10.1007%2F978-0-387-30164-8). ISBN 978-0-387-30164-8.
6. Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research" (https://www.cawcr.gov.au/projects/verification/). *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
7. Chicco D.; Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).
8. Chicco D.; Toetsch N.; Jurman G. (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). *BioData Mining*. **14** (13): 13. doi:10.1186/s13040-021-00244-z (https://doi.org/10.1186%2Fs13040-021-00244-z). PMC 7863449 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). PMID 33541410 (https://pubmed.ncbi.nlm.nih.gov/33541410).
9. Chicco D.; Jurman G. (2023). "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9938573). *BioData Mining*. **16** (1): 4. doi:10.1186/s13040-023-00322-4 (https://doi.org/10.1186%2Fs13040-023-00322-4). PMC 9938573 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9938573). PMID 36800973 (https://pubmed.ncbi.nlm.nih.gov/36800973).
10. Tharwat A. (August 2018). "Classification assessment methods" (https://doi.org/10.1016%2Fj.aci.2018.08.003). *Applied Computing and Informatics*. **17**: 168–192. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).
11. Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7540853). *PLOS ONE*. **15** (10): e0240215. arXiv:2006.00398 (https://arxiv.org/abs/2006.00398). Bibcode:2020PLoSO..1540215B (https://ui.adsabs.harvard.edu/abs/2020PLoSO..1540215B). doi:10.1371/journal.pone.0240215 (https://doi.org/10.1371%2Fjournal.pone.0240215). PMC 7540853 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7540853). PMID 33027310 (https://pubmed.ncbi.nlm.nih.gov/33027310).
12. Saah AJ, Hoover DR (1998). "[Sensitivity and specificity revisited: significance of the terms in analytic and diagnostic language]" (https://pubmed.ncbi.nlm.nih.gov/9747274). *Ann Dermatol Venereol*. **125** (4): 291–4. PMID 9747274 (https://pubmed.ncbi.nlm.nih.gov/9747274).
13. Parikh, Rajul; Mathai, Annie; Parikh, Shefali; Chandra Sekhar, G; Thomas, Ravi (2008). "Understanding and using sensitivity, specificity and predictive values" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636062). *Indian Journal of Ophthalmology*. **56** (1): 45–50. doi:10.4103/0301-4738.37595 (https://doi.org/10.4103%2F0301-4738.37595). PMC 2636062 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636062). PMID 18158403 (https://pubmed.ncbi.nlm.nih.gov/18158403).
14. Altman DG, Bland JM (June 1994). "Diagnostic tests. 1: Sensitivity and specificity" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2540489). *BMJ*. **308** (6943): 1552. doi:10.1136/bmj.308.6943.1552 (https://doi.org/10.1136%2Fbmj.308.6943.1552). PMC 2540489 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2540489). PMID 8019315 (https://pubmed.ncbi.nlm.nih.gov/8019315).
15. "SpPin and SnNout" (https://www.cebm.ox.ac.uk/resources/ebm-tools/sppin-and-snnout). Centre for Evidence Based Medicine (CEBM). Retrieved 18 January 2023.
16. Mangrulkar R. "Diagnostic Reasoning I and II" (http://open.umich.edu/education/med/m1/patientspop-decisionmaking/2010/materials). Retrieved 24 January 2012.
17. "Evidence-Based Diagnosis" (https://web.archive.org/web/20130706035232/http://omerad.msu.edu/ebm/Diagnosis/Diagnosis4.html). Michigan State University. Archived from the original (http://omerad.msu.edu/ebm/Diagnosis/Diagnosis4.html) on 2013-07-06. Retrieved 2013-08-23.
18. "Sensitivity and Specificity" (http://www.med.emory.edu/EMAC/curriculum/diagnosis/sensand.htm). Emory University Medical School Evidence Based Medicine course.
19. Baron JA (Apr–Jun 1994). "Too bad it isn't true". *Medical Decision Making*. **14** (2): 107. doi:10.1177/0272989X9401400202 (https://doi.org/10.1177%2F0272989X9401400202). PMID 8028462 (https://pubmed.ncbi.nlm.nih.gov/8028462). S2CID 44505648 (https://api.semanticscholar.org/CorpusID:44505648).
20. Boyko EJ (Apr–Jun 1994). "Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn?". *Medical Decision Making*. **14** (2): 175–9. doi:10.1177/0272989X9401400210 (https://doi.org/10.1177%2F0272989X9401400210). PMID 8028470 (https://pubmed.ncbi.nlm.nih.gov/8028470). S2CID 31400167 (https://api.semanticscholar.org/CorpusID:31400167).
21. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M (July 2004). "Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC487735). *BMJ*. **329** (7459): 209–13. doi:10.1136/bmj.329.7459.209 (https://doi.org/10.1136%2Fbmj.329.7459.209). PMC 487735 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC487735). PMID 15271832 (https://pubmed.ncbi.nlm.nih.gov/15271832).
22. Fawcett, Tom (2006). "An Introduction to ROC Analysis". *Pattern Recognition Letters*. **27** (8): 861–874. Bibcode:2006PaReL..27..861F (https://ui.adsabs.harvard.edu/abs/2006PaReL..27..861F). CiteSeerX 10.1.1.646.2144 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.646.2144). doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010). S2CID 2027090 (https://api.semanticscholar.org/CorpusID:2027090).
23. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
24. Gale SD, Perkel DJ (January 2010). "A basal ganglia pathway drives selective auditory responses in songbird dopaminergic neurons via disinhibition" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2824341). *The Journal of Neuroscience*. **30** (3): 1027–37. doi:10.1523/JNEUROSCI.3585-09.2010 (https://doi.org/10.1523%2FJNEUROSCI.3585-09.2010). PMC 2824341 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2824341). PMID 20089911 (https://pubmed.ncbi.nlm.nih.gov/20089911).
25. Macmillan NA, Creelman CD (15 September 2004). ***Detection Theory: A User's Guide*** (https://books.google.com/books?id=hDX65v9bReYC). Psychology Press. p. 7. ISBN 978-1-4106-1114-7.
26. Balayla, Jacques (2020). "Prevalence threshold (ϕe) and the geometry of screening curves" (https://doi.org/10.1371%2Fjournal.pone.0240215). *PLOS ONE*. **15** (10): e0240215. doi:10.1371/journal.pone.0240215 (https://doi.org/10.1371%2Fjournal.pone.0240215). PMID 33027310 (https://pubmed.ncbi.nlm.nih.gov/33027310).
27. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010). S2CID 2027090 (https://api.semanticscholar.org/CorpusID:2027090).
28. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512 (https://doi.org/10.1061%2F%28ASCE%29IS.1943-555X.0000512). S2CID 213782055 (https://api.semanticscholar.org/CorpusID:213782055).
29. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
30. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8 (https://doi.org/10.1007%2F978-0-387-30164-8). ISBN 978-0-387-30164-8.
31. Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research" (https://www.cawcr.gov.au/projects/verification/). *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
32. Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).
33. Chicco D, Toetsch N, Jurman G (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). *BioData Mining*. **14** (13): 13. doi:10.1186/s13040-021-00244-z (https://doi.org/10.1186%2Fs13040-021-00244-z). PMC 7863449 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). PMID 33541410 (https://pubmed.ncbi.nlm.nih.gov/33541410).
34. Tharwat A. (August 2018). "Classification assessment methods" (https://doi.org/10.1016%2Fj.aci.2018.08.003). *Applied Computing and Informatics*. **17**: 168–192. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).
35. Lin, Jennifer S.; Piper, Margaret A.; Perdue, Leslie A.; Rutter, Carolyn M.; Webber, Elizabeth M.; O'Connor, Elizabeth; Smith, Ning; Whitlock, Evelyn P. (21 June 2016). "Screening for Colorectal Cancer" (https://doi.org/10.1001/jama.2016.3332). *JAMA*. **315** (23): 2576–2594. doi:10.1001/jama.2016.3332 (https://doi.org/10.1001%2Fjama.2016.3332). ISSN 0098-7484 (https://www.worldcat.org/issn/0098-7484). PMID 27305422 (https://pubmed.ncbi.nlm.nih.gov/27305422).
36. Bénard, Florence; Barkun, Alan N.; Martel, Myriam; Renteln, Daniel von (7 January 2018). "Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5757117). *World Journal of Gastroenterology*. **24** (1): 124–138. doi:10.3748/wjg.v24.i1.124 (https://doi.org/10.3748%2Fwjg.v24.i1.124). PMC 5757117 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5757117). PMID 29358889 (https://pubmed.ncbi.nlm.nih.gov/29358889).
37. "Diagnostic test online calculator calculates sensitivity, specificity, likelihood ratios and predictive values from a 2x2 table – calculator of confidence intervals for predictive parameters" (http://www.medcalc.org/calc/diagnostic_test.php). *medcalc.org*.

Further reading

- Altman DG, Bland JM (June 1994). "Diagnostic tests. 1: Sensitivity and specificity" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2540489). *BMJ*. **308** (6943): 1552. doi:10.1136/bmj.308.6943.1552 (https://doi.org/10.1136%2Fbmj.308.6943.1552). PMC 2540489 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2540489). PMID 8019315 (https://pubmed.ncbi.nlm.nih.gov/8019315).
- Loong TW (September 2003). "Understanding sensitivity and specificity with the right side of the brain" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC200804). *BMJ*. **327** (7417): 716–9. doi:10.1136/bmj.327.7417.716 (https://doi.org/10.1136%2Fbmj.327.7417.716). PMC 200804 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC200804). PMID 14512479 (https://pubmed.ncbi.nlm.nih.gov/14512479).

External links

- UIC Calculator (http://araw.mede.uic.edu/cgi-bin/testcalc.pl)
- Vassar College's Sensitivity/Specificity Calculator (http://vassarstats.net/clin1.html)
- MedCalc Free Online Calculator (https://www.medcalc.org/calc/diagnostic_test.php)
- Bayesian clinical diagnostic model applet (https://kennis-research.shinyapps.io/Bayes-App/)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Sensitivity_and_specificity&oldid=1198560681"