Sensitivity and specificity

From Wikipedia, the free encyclopedia

Sensitivity and **specificity** are statistical measures of the performance of a binary classification test, also known in statistics as classification function. **Sensitivity** (also called the *true positive rate*, or the **recall rate** in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition). **Specificity** measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition, sometimes called the *true negative rate*). These two measures are closely related to the concepts of type I and type II errors. A perfect predictor would be described as 100% sensitive (i.e. predicting all people from the sick group as sick) and 100% specific (i.e. not predicting anyone from the healthy group as sick); however, theoretically any predictor will possess a minimum error bound known as the Bayes error rate.

For any test, there is usually a trade-off between the measures. For example: in an airport security setting in which one is testing for potential threats to safety, scanners may be set to trigger on low-risk items like belt buckles and keys (low specificity), in order to reduce the risk of missing objects that do pose a threat to the aircraft and those aboard (high sensitivity). This trade-off can be represented graphically as a receiver operating characteristic curve.

Contents [hide]

Definitions [edit]

Imagine a study evaluating a new 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 (predicting that the person has the disease) or negative (predicting that the person does not have 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 diagnosed 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

In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

Sensitivity [edit]

Sensitivity relates to the test's ability to identify positive results.

The sensitivity of a test is the proportion of people that are known to have the disease who test positive for it. This can also be written as:

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}}$$

$$= \text{probability of a positive test, given that the patient is ill}$$

Again, consider the example of the medical test used to identify a disease. A 'bogus' test kit that always indicates positive regardless of the disease status of the patient will achieve, from a theoretical point of view, 100% sensitivity. This is because in this case there are no negatives at all, and false positives are not accounted for in the definition of sensitivity. Therefore, sensitivity alone cannot be used to determine whether a test is useful in practice.

However, a test with high sensitivity can be considered as a reliable indicator when its result is negative, since it rarely misses true positives among those who are actually positive. For example, a sensitivity of 100% means that the test recognizes all actual positives – i.e. all sick people are recognized as being ill. Thus, in contrast to a high specificity test, *negative results* in a *high sensitivity test* are used to *rule out* the disease.

Sensitivity is not the same as the precision or positive predictive value (ratio of true positives to combined true and false positives), which is as much a statement about the proportion of actual positives in the population being tested as it is about the

test.

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 high sensitivity has a low type II error rate. In non-medical contexts, sensitivity is sometimes called recall.

Specificity [edit]

Specificity relates to the test's ability to identify negative results.

Consider the example of the medical test used to identify a disease. The specificity of a test is defined as the proportion of patients that are known not to have the disease who will test negative for it. This can also be written as:

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

$$= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}}$$

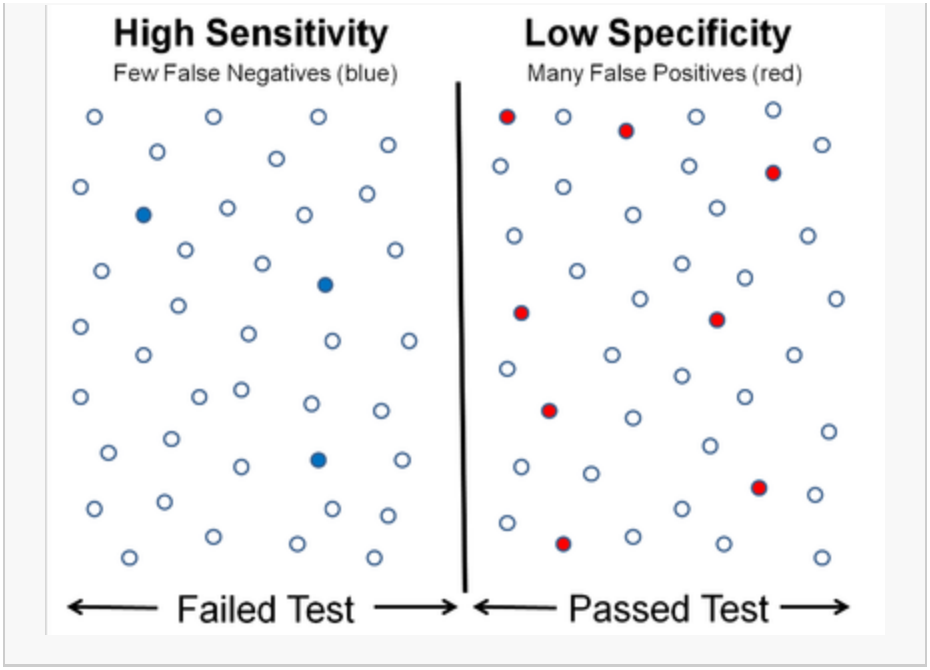
$$= \text{probability of a negative test given that the patient is well}$$

From a theoretical point of view, a 'bogus' test kit which always indicates *negative* regardless of the disease status of the patient, will achieve 100% specificity, since there are no positive results and false negatives are not accounted for by definition.

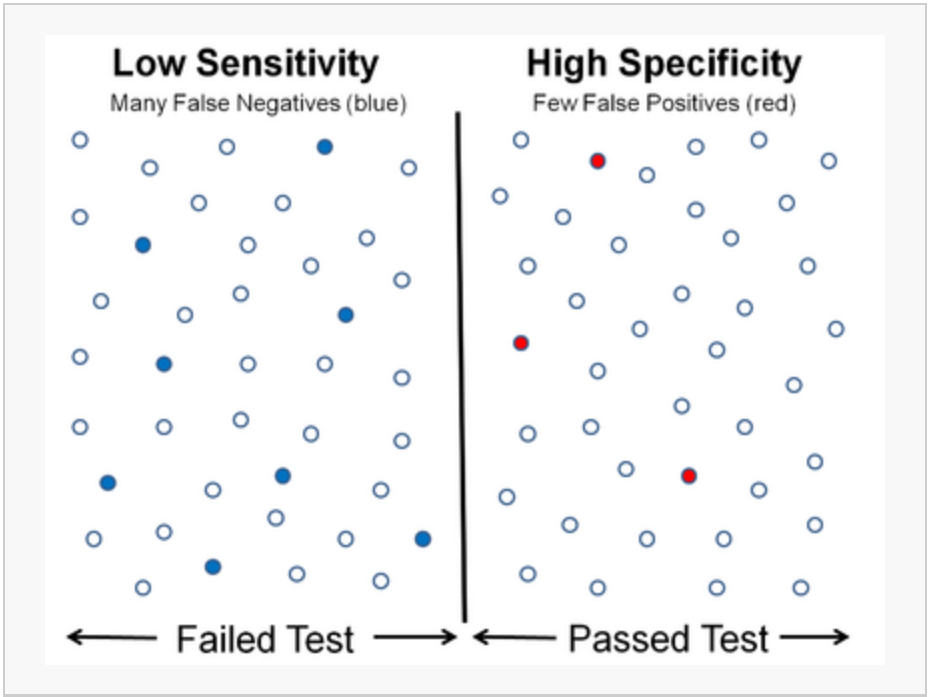
However, highly specific tests rarely miss negative outcomes, so they can be considered reliable when their result is *positive*. Therefore, a positive result from a test with high specificity means a high probability of the presence of disease.[1]

A test with a high specificity has a low type I error rate.

Graphical illustration [edit]



High sensitivity and low specificity



Low sensitivity and high specificity

Medical examples [edit]

In medical diagnostics, 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 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.[2] 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.

Misconceptions [edit]

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.[3][4] This has led to the widely used mnemonics SPIN and SNOUT, according to which a highly SPecific test, when Positive, rules IN disease (SP-P-IN), and a highly 'SeNsitive' test, when Negative rules OUT disease (SN-N-OUT). Both rules of thumb are, howevever, inferentially misleading, as the diagnostic power of any test is determined by *both* the sensitivity and specificity.[5][6][7]

Worked example [edit]

Relationships among terms

view · talk · edit ·

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

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%

Patients with bowel cancer (as confirmed on endoscopy)

		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) **≈ 99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) **≈ 67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

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
- Likelihood ratio positive = sensitivity / (1 − specificity) = 66.67% / (1 − 91%) = 7.4
- Likelihood ratio negative = (1 − sensitivity) / specificity = (1 − 66.67%) / 91% = 0.37

Hence with large numbers of false positives and few false negatives, a positive screen test is in itself poor at confirming the disorder (PPV = 10%) and further investigations must be undertaken; it did, however, correctly identify 66.7% of all cases (the sensitivity). However as a screening test, a negative result is very good at reassuring that a patient does not have the disorder (NPV = 99.5%) and at this initial screen correctly identifies 91% of those who do not have cancer (the specificity).

Estimation of errors in quoted sensitivity or specificity [edit]

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%).[8]

Terminology in information retrieval [edit]

In information retrieval, the positive predictive value is called **precision**, and sensitivity is called **recall**.

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

recall:

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

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 [edit]

- Accuracy and precision
- Brier score
- Confusion matrix
- Detection theory
- F-score
- Gain (information retrieval)
- Likelihood ratios
- Matthews correlation coefficient
- Receiver operating characteristic or ROC curve
- Selectivity
- Sensitivity index
- Statistical significance
- Youden's J statistic
- OpenEpi software program

References [edit]

1. ^ http://www.cebm.net/index.aspx?o=1042
2. ^ Mangrulkar, Rajesh. "Diagnostic Reasoning I and II". Retrieved 24 January 2012.
3. ^ Michigan State University Evidence Based Medicine resource
4. ^ Emory University Medical School Evidence Based Medicine course
5. ^ Baron, J A Too Bad It Isn't True *Med Decis Making* 1994 **14**: 107
6. ^ Boyko, E J Ruling Out or Ruling In Disease with the Most sensitive or specific diagnostic test: short cut or wrong turn ? *Med Decis Making* 1994 **14** 175-179
7. ^ Pewsner, D et al, Ruling a diagnosis in or out with SpPIn and SnNOut: a note of caution *British Medical Journal*

Terminology and derivations from a confusion matrix

true positive (TP)
eqv. with hit

true negative (TN)
eqv. with correct rejection

false positive (FP)
eqv. with false alarm, Type I error

false negative (FN)
eqv. with miss, Type II error

sensitivity or true positive rate (TPR)
eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

false positive rate (FPR)
eqv. with false alarm rate, fall-out

$$FPR = FP/N = FP/(FP + TN)$$

accuracy (ACC)

$$ACC = (TP + TN)/(TP + TN + FP + FN)$$

specificity (SPC) or True Negative Rate

$$SPC = TN/N = TN/(FP + TN) = 1 - FPR$$

positive predictive value (PPV)
eqv. with precision

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP)$$

Matthews correlation coefficient (MCC)

2004 **329** 209

8. ^ Online calculator of confidence intervals for predictive parameters

$$MCC = (TPTN - FPFN)/\sqrt{PNP'N'}$$

Source: Fawcett (2004).

Further reading [edit]

- Altman DG, Bland JM (1994). "Diagnostic tests. 1: Sensitivity and specificity". *BMJ* **308** (6943): 1552. PMC 2540489. PMID 8019315.
- Loong T (2003). "Understanding sensitivity and specificity with the right side of the brain". *BMJ* **327** (7417): 716–719. doi:10.1136/bmj.327.7417.716. PMC 200804. PMID 14512479.

External links [edit]

- Vassar College's Sensitivity/Specificity Calculator

v · t · e ·	**Biomedical research: Clinical study design / Design of experiments** [hide]
Overview	Clinical trial (Trial protocols · · Academic clinical trials · Clinical study design ·
Controlled study (EBM I to II-1; A to B)	Randomized controlled trial (Blind experiment, Open-label trial) ·
Observational study (EBM II-2 to II-3; B to C)	Cross-sectional study vs. Longitudinal study, Ecological study · Cohort study (Retrospective · Prospective · · Case-control study (Nested case-control study) · Case series · Case study · Case report ·
Epidemiology/ methods	*occurrence:* Incidence (Cumulative incidence) · Prevalence (Point · Period · · *association: absolute* (Absolute risk reduction, Attributable risk, Attributable risk percent) · *relative* (Relative risk, Odds ratio, Hazard ratio) · *other:* End point of clinical trials · Virulence · Infectivity · Mortality rate · Morbidity · Case fatality rate · Specificity and sensitivity · Likelihood-ratios · Pre/post-test probability ·
Trial/test types	In vitro · In vivo · Animal testing · Animal testing on non-human primates · First-in-man study · Multicenter trial · Seeding trial · Vaccine trial ·
Analysis of clinical trials	Risk–benefit analysis · Systematic review · Meta-analysis ·
Interpretation of results	Selection bias · Correlation does not imply causation · Null result ·
	Category · Glossary · List of topics ·

Categories: Statistical theory | Biostatistics | Medical statistics | Statistical ratios | Bioinformatics | Cheminformatics | Summary statistics for contingency tables