# Statistical **Measurements in Medical Ultrasound**

DIAGNOSTIC PERFORMANCE | CLINICAL CONFIDENCE

**2 x 2 DIAGNOSTIC CONTINGENCY TABLE**

| | |
|---|---|
| TRUE POSITIVE **TP** | FALSE POSITIVE **FP** |
| FALSE NEGATIVE **FN** | TRUE NEGATIVE **TN** |

# Statistical Measurements in Medical Ultrasound

## Why Statistics?

Modern ultrasound is both a visual imaging modality and a diagnostic test. Because sonographic findings influence clinical decisions, test performance should be described using **objective statistical measures** rather than **subjective impressions** such as "looks reliable." Statistical evaluation in medical testing is used to quantify how trustworthy a test is, how often it is correct, and how much confidence should be placed in a positive or negative result, providing evidence-based guidance for safe clinical decision-making. Integrating statistical methods is associated with benefits to both patient and practitioner.

***Benefits to patient:***

- Fewer missed diagnoses
- Fewer unnecessary procedures & anxiety
- More appropriate triage and care

***Benefits to practitioner:***

- Clearer clinical confidence in results
- Improved quality assurance and performance monitoring
- Smarter election of the right test

## Four Core Diagnostic Parameters

**Sensitivity, specificity, positive predictive value (*PPV*), and negative predictive value (*NPV*)** are four core parameters used to describe how well a diagnostic test performs and how confidently its results can be applied in clinical care. **Sensitivity** and **specificity** summarize test behavior under known truth conditions, how effectively disease is detected and how reliably non-disease is excluded, while **PPV** and **NPV** translate test results into practical clinical meaning by estimating how trustworthy a positive or negative result is in the population being tested. Together, these measures provide a standardized way to interpret ultrasound findings, communicate diagnostic confidence, and reduce both missed disease and unnecessary follow-up.

- **Sensitivity**: how often a test ***identifies*** disease when disease is truly present
- **Specificity**: how well a test ***excludes*** disease when disease is truly absent
- **PPV**: the probability disease is truly ***present when*** test result is ***positive***
- **NPV**: the probability disease is truly ***absent when*** test result is ***negative***

## The Foundation: 2x2 Contingency Table

The 2×2 contingency table is the basic framework used to evaluate any diagnostic test, including ultrasound. It compares what the test reports (***positive*** *or* ***negative***) with what is actually true based on a reference standard (***disease present*** *or* ***disease absent***). Each patient outcome is placed into one of four categories, sometimes called the four outcomes of a diagnostic test:

- **TP (*True Positive*):** the test is positive, and disease is truly present
- **FP (*False Positive*):** the test is positive, but disease is truly absent
- **FN (*False Negative*):** the test is negative, but disease is truly present
- **TN (*True Negative*):** the test is negative, and disease is truly absent

By converting clinical results into these simple counts (*TP, FP, FN, TN*), the contingency table provides the values needed to calculate key diagnostic performance parameters such as **sensitivity, specificity, positive predictive value (*PPV*),** and **negative predictive value (*NPV*)**.

The table is especially useful for calculating **sensitivity** and **specificity** because it separates patients into groups based on the reference standard. **Sensitivity** is calculated from the **disease-present** group and describes the proportion of truly diseased patients who test positive:

$$\textbf{Sensitivity} = TP \div (TP + FN)$$

**Specificity** is calculated from the **disease-absent** group and describes the proportion of truly non-diseased patients who test negative:

$$\textbf{Specificity} = TN \div (TN + FP)$$

In practical terms, the 2×2 contingency table standardizes how test outcomes are organized, making it easier to describe diagnostic reliability and to translate ultrasound results into clinical confidence.

| 2x2 CONTINGENCY TABLE | | |
|---|---|---|
| | **DISEASE PRESENT** | **DISEASE ABSENT** |
| **TEST POSITIVE** | TP (*true positive*) | FP (*false positive*) |
| **TESTE NEGATIVE** | FN (*false negative*) | TN (*true negative*) |

**Values Calculated from Contingency Table**

**Sensitivity**

*"How good is the test at finding disease?"*

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

**Specificity**

*"How good is the test at excluding disease?"*

$$\text{Specificity} = \frac{TN}{TN + FP}$$

**Postive Predictive Value (PPV)**

*"If the test is positive, how likely is true disease?"*

$$PPV = \frac{TP}{TP + FP}$$

**Negative Predictive Value (NPV)**

*"If the test is negative, how likely is truly no disease?"*

$$NPV = \frac{TN}{TN + FN}$$

**Example Calculations:**

**Calculate the diagnostic parameters for patients who are referred for a lower extremity venous ultrasound exam to rule-out deep vein thrombosis (r/o DVT)**

- **Sensitivity** – How often ultrasound correctly identifies disease when disease is present.

  Calculate the sensitivity given the following data:
  - 100 patients have DVT
  - Ultrasound correctly reports 92 positive results (***TP***)
  - Ultrasound incorrectly reports (*misses*) 8 as negative fort DVT (***FN***)

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{92}{92 + 8} = \frac{92}{100}$$

$$\text{Sensitivity} = 92\%$$

- **Specificity** – How often ultrasound correctly identifies no disease when disease is absent. Describes how well an ultrasound test can correctly identify patients without the condition being tested for.

  Calculate the specificity given the following data:
  - 100 patients do **NOT** have DVT
  - Ultrasound correctly reports that 92 do NOT have DVT(***TN***)
  - Ultrasound incorrectly reports (*misses*) 8 as negative fort DVT (***FP***)

$$\text{Specificity} = \frac{TN}{TN + FP}$$
$$= \frac{92}{92 + 8}$$
$$= \frac{92}{100}$$
$$\text{Specificity} = 92\%$$

- **Positive Predictive Value -** The probability that a positive ultrasound result truly indicates disease
  - 150 patients referred to r/o DVT
  - 100 patients truly have DVT
    - US correctly identifies 92 as positive (***TP***)
    - US misses 8 and reports as negative (***FN***)
  - 50 patients do NOT have DVT
    - US correctly reports 40 as negative (***TN***)
    - US incorrectly reports 10 as positive (***FP***)

$$PPV = \frac{TP}{TP + FP}$$
$$= \frac{92}{92 + 10}$$
$$= \frac{92}{102}$$
$$PPV = 90.2\%$$

- **Negative Predictive Value -** The probability that a negative ultrasound result truly indicates absence of disease.
  - 150 patients referred to r/o DVT
  - 100 patients truly have DVT
    - US correctly identifies 92 as positive (***TP***)
    - US misses 8 and reports as negative (***FN***)

- 50 patients do NOT have DVT
  - US correctly reports 40 as negative (***TN***)
  - US incorrectly reports 10 as positive (***FP***)

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} = \frac{40}{40 + 8} = \frac{40}{48}$$
$$\text{NPV} = 83.3\%$$

## Interpreting Clinical Results

Clinical interpretation of diagnostic statistics goes beyond calculating sensitivity, specificity, PPV, and NPV. The practical goal is to decide what a test result *means* for a specific patient and how strongly it should influence management. The same ultrasound result can carry different clinical weight depending on the test's performance characteristics, the patient's symptoms, and how likely disease was before imaging was performed (*pretest probability*). This section explains three core interpretation skills: using "rule out" versus "rule in" logic, understanding how prevalence changes predictive values, and recognizing why a negative study does not always end the workup when clinical suspicion remains high.

- "***Rule out***" vs "***rule in***" logic (*SnNout* / *SpPin*)
  Two common memory cues summarize how sensitivity and specificity guide clinical thinking:
  - ***SnNout:*** A highly SeNsitive test, when Negative, helps rule OUT disease. This works because a very sensitive test produces few false negatives (*FN*). If disease were truly present, the test would usually be positive; therefore, a negative result makes disease less likely.
  - ***SpPin***: A highly SPecific test, when Positive, helps rule IN disease. This works because a very specific test produces few false positives (*FP*). If the test is positive, it is unlikely to be a mistaken positive; therefore, a positive result strongly supports disease.

  These rules are not absolute. They simply describe where the greatest confidence tends to lie when sensitivity or specificity is very high.

- **How prevalence changes PPV and NPV**

  **PPV and NPV depend on prevalence** (*how common the disease is in the tested population*). Sensitivity and specificity describe test behavior, but PPV/NPV describe how trustworthy the result is in the setting where the test is being used. A simple way to remember the direction of change:

  - When **prevalence increases**, **PPV increases** and **NPV decreases**.
  - When **prevalence decreases**, **PPV decreases** and **NPV increases**.

- **What a negative study means when suspicion is high (*false negatives*)**

  A negative test result reduces the probability of disease, but it does not always eliminate it. When clinical suspicion is high, a negative study may represent a **false negative**, a case in which disease is present but not detected. False negatives occur for several reasons:

  - Disease may be early, small, subtle, or intermittent
  - Patient factors may limit image quality (*body habitus, bowel gas, poor acoustic windows*)
  - Technical factors may limit detection (*suboptimal settings, incomplete visualization, operator dependence*)
  - The suspected diagnosis may involve a related condition that the specific test is less sensitive for

  In practice, a negative ultrasound in a high-suspicion patient should trigger a second question: *Is the test sensitive enough in this scenario to confidently exclude disease?* If not, appropriate next steps may include repeat imaging, an alternative modality, laboratory correlation, or follow-up based on clinical judgment. The key clinical takeaway is that test results are interpreted most safely when statistics are combined with the clinical picture rather than used in isolation.

## Chi-square in Imaging

Chi-square ($\chi^2$) is a statistical test used to determine whether there is a real association between two categorical variables, or whether an observed difference in counts is likely due to random variation. In diagnostic imaging, many outcomes are recorded as **categories** rather than continuous measurements. Examples include *test positive vs. test negative*, *disease present vs. disease absent*, *pass v. fail on quality assurance checks*, or *artifact present vs. absent*. When results are expressed as counts in categories, the **chi-square ($\chi^2$)** test provides a standardized method for determining whether an observed pattern is likely to represent a real effect or whether it could be explained by random variation. In simple terms, chi-square helps answer the question:

***"Do the numbers differ more than would be expected by chance?"***

**The core idea:** Chi-square compares:

- Observed counts ($O$) = what was actually measured

***vs***

- Expected counts ($E$) = what would be expected if there were no real difference or no association

The chi-square statistic is: $$\chi^2 = \sum \frac{(O-E)^2}{E}$$

A large $\chi^2$ value means the observed pattern differs strongly from what would be expected by chance. Output is probability (***p***) of seeing whether results are just by random chance or if there is a statistically valid association.

Chi-square analysis is commonly introduced in two practical imaging contexts. First, it can evaluate whether **two categorical variables are related**, such as whether an ultrasound result (*positive/negative*) is associated with true disease status (*present/absent*). Second, it can determine whether the **distribution of imaging outcomes** matches an expected pattern, such as whether a lab's gallbladder ultrasound findings this month (*normal, stones, cholecystitis, polyps*) resemble the department's typical historical proportions. These two uses are presented below as (1) a test of independence and (2) a goodness-of-fit test.

- **Chi-square test of independence (***association between two variables***)**
  The test of independence is used when there are **two categorical variables**, often displayed as a contingency table (*commonly 2×2 in medical testing*). It evaluates whether the variables are statistically associated.

  ***Imaging example:***
  Variable 1: ultrasound result (*positive vs negative*)
  Variable 2: true disease status (*present vs absent*)

  ***Meaning of the result:***
  A small p-value (*e.g., $p < 0.05$*) suggests ultrasound results and disease status are not independent; they are related in a way unlikely due to chance.
  A large p-value suggests there is not enough evidence to conclude a relationship.

- **Chi-square goodness-of-fit (***does one variable match an expected distribution?***)**
  The goodness-of-fit test is used when there is **one categorical variable**, and the goal is to compare observed counts to a known or expected distribution.

**Imaging example:**
A department expects gallbladder ultrasound outcomes to be approximately: 50% normal, 30% stones, 15% cholecystitis, 5% polyps.

This month's observed case counts can be tested against those expected proportions. If the chi-square result is significant, the observed distribution **does not fit** the expected pattern—suggesting a meaningful shift in referral patterns, patient population, technique, or reporting practices.

**Practical "remember that…"**

Chi-square is most useful when imaging data are **counts in categories**. It does not measure "how good" a test is by itself; rather, it evaluates whether observed differences or distributions are **likely real** rather than random.

## Standard Statistical Metrics

### Core numerical measures

- **Mean** – The average value of a set of measurements
- **Median** – The middle value; useful when a few extreme values might distort the mean
- **Range** – The spread from the smallest to the largest value
- **Standard deviation (*SD*)** – Describes how much values typically vary around the mean

### Comparing variability

- **Variance** – The average squared distance of values from the mean (*basis for SD*)
- **Coefficient of variation (*CV*)** – SD divided by the mean (*often as a percentage*); useful for comparing variability between different measurements, operators, or systems