Probability Reversal

When a test’s accuracy is read backwards

Murray Cantor PhD

Fault 2 of six · healthy-skepticism.com · September 2026

The fault

A test result arrives with a number attached: "This test is 96% accurate." A positive comes back, and the number seems to settle it. Patients act on that reading. They start the treatment, isolate from the family, or lose a week to worry, when the right next step was a confirming test. A negative from the same test is read as the all-clear, and acted on too.

The number is real, and the reading is wrong. The fault is probability reversal. The accuracy figure answers a question about the test. The patient is asking a question about herself, and the answers can differ wildly. A positive from an accurate test can still mean the condition is unlikely; a negative can still leave it in play. The example the course uses: does a positive COVID test mean I have it? Does a negative mean I'm not contagious? A healthy skeptic reads the result as evidence, not a verdict. She neither dismisses it nor acts on it as if it were certain. She asks what it actually says about her, confirms the test with perhaps a scan or a biopsy, and holds off on treatment or relief until she has the answer. This paper explains why the number on the box isn't the answer, what else is needed, and how to ask for it.

Conditional probability

Nothing in medicine is certain. Every diagnosis, every treatment choice, and every published trial result is a statement about how likely something is, given what is known: the probability of a disease given a symptom, of a heart attack given a drug, of recovery given a treatment. The chance of one thing given that another is true is called a conditional probability, and it is the underlying mechanism of all medical research and medical decision making. When a doctor orders a test, she is asking for evidence that will change a conditional probability. When she reads the result, she is updating one.

It is also how we make decisions everywhere else, whether or not we use the words. A jury is instructed to convict only when guilt is proven "beyond a reasonable doubt." Stripped of the legal phrasing, the instruction reads: given the evidence presented, is the probability of guilt high enough to decide the defendant is guilty? The evidence doesn't prove guilt. It raises or lowers the probability of guilt, and the standard of proof is a threshold on that probability. A juror deciding whether to convict, a doctor deciding whether to order a biopsy, and a patient deciding whether to take a drug are all doing the same thing: weighing a probability given the evidence against the cost of being wrong.

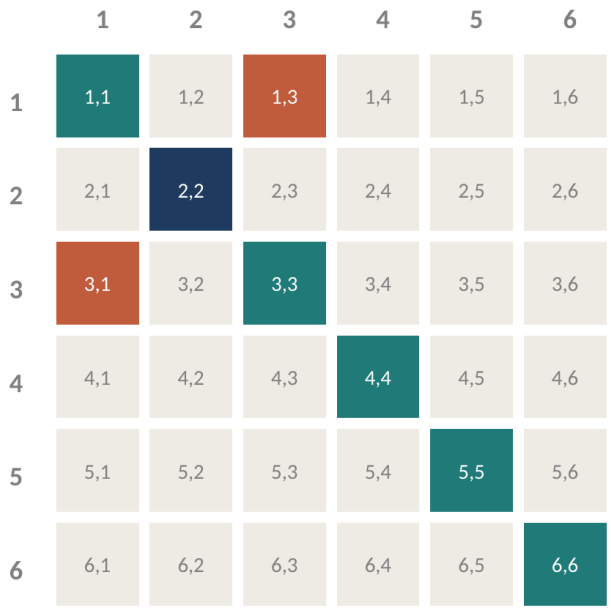
Because conditional probability is the mechanism, its faults are the faults of medical reasoning, and reading one in the wrong direction is the most common of them. Seeing exactly why takes a little notation and two dice.

The notation, with two dice

Write P(X | Y) for the probability of X given Y, read “the probability of X, given that Y is true.” To find it, keep only the cases where Y holds, then count how many of those are X. Two ordinary dice make the

whole idea visible. There are 36 equally likely rolls. Three of them add to four, (1,3), (2,2), and (3,1), so P(roll = 4) is 3 of 36. Six are doubles, (1,1) through (6,6), so P(double) is 6 of 36.

Now ask the two conditional questions. Given that the roll adds to four, is it a double? Keep the three fours; one of them, (2,2), is a double: P(double | roll = 4) is 1 in 3. Given that the roll is a double, does it add to four? Keep the six doubles; one of them, again (2,2), adds to four: P(roll = 4 | double) is 1 in 6. It is the same square both times. The two probabilities differ because each counts that square against a different group, the three fours or the six doubles. That is the whole of probability reversal: same evidence, different denominator, different answer.

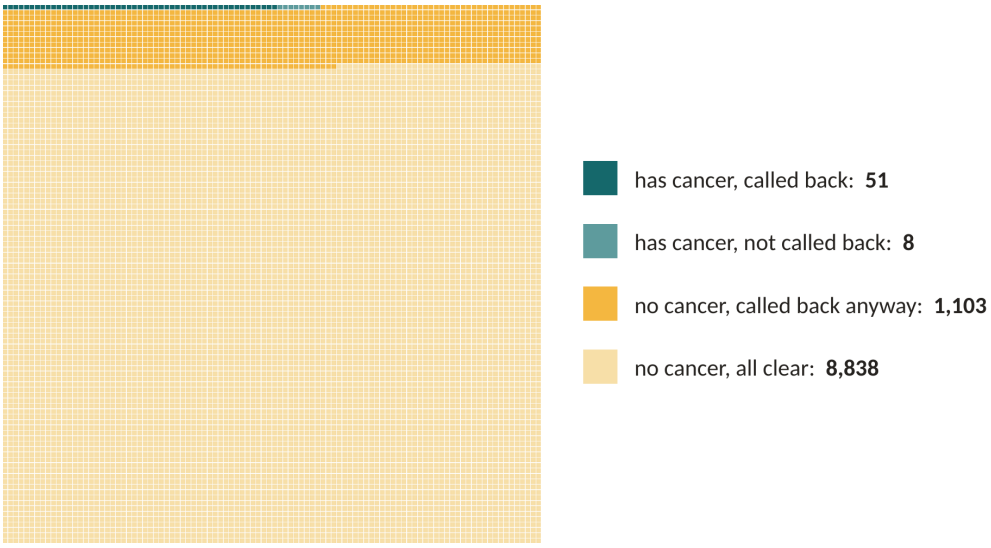


The 36 rolls. Terracotta adds to four, teal is a double, and (2,2), in navy, is both.

P(double | four) is 1 in 3. P(four | double) is 1 in 6. Same square, counted against two different groups.

The example: a mammogram callback

Screening mammography detects about 87% of the cancers present. Among women being screened, roughly 6 in 1,000 have cancer. So out of 1,000 women screened, about 6 have cancer, and the test calls back 5 of them. It also calls back about 110 of the 994 women who do not have cancer, because callbacks are common and cancer is not. A woman who receives the letter is one of about 115, and 5 of those 115 have cancer. The chance that her callback means cancer is about 1 in 23, roughly 4%.



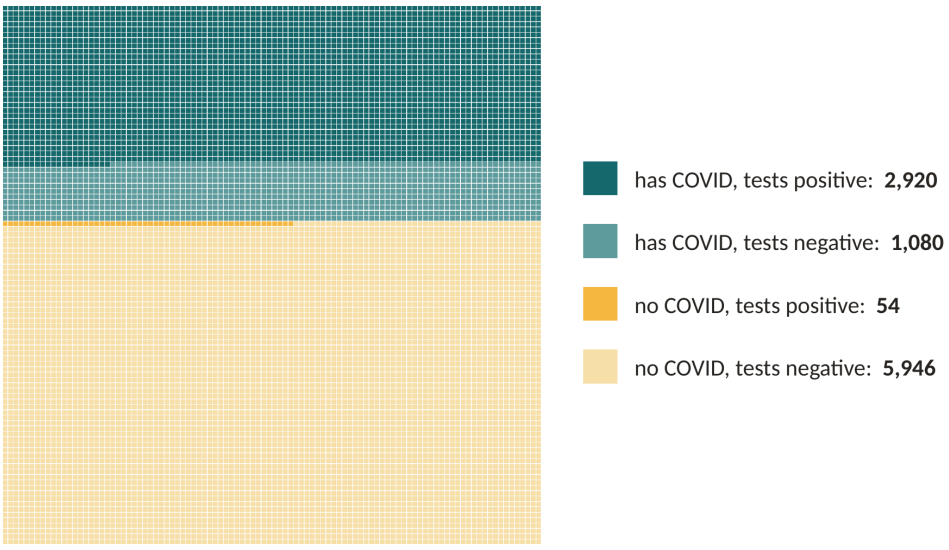
Ten thousand women screened once. Of the 1,154 called back, 51 have cancer.

Same test, same day, same woman: 87% in the measured direction, 4% in the direction she cares about. Unlikely is why she should not panic.

Lehman CD et al. National performance benchmarks for modern screening digital mammography. Radiology 2017;283:49–58. Figures rounded to whole women per 1,000 for teaching.

The COVID tests, and the engineering behind the asymmetry

The pandemic put two tests for one virus into every household, and they were engineered in opposite directions. The rapid antigen test was built for specificity. In symptomatic people it correctly clears about 99 of every 100 who are not infected, and that is the number it is sold on. Its sensitivity is another matter: it catches about 73 of every 100 who are infected, which means about 27 walk out with a negative result and the infection. A positive rapid test is strong evidence. A negative one is not the all-clear, and the instructions say so: retest.



Ten thousand people with symptoms during a wave, four in ten infected. The 1,080 who have COVID and test negative are the band to watch.

PCR was built the other way, for sensitivity. It was tuned to detect minute traces of viral RNA so as to miss no one, and the price appears in the other direction: positives that persist for weeks after a person has stopped being infectious. For the question a patient actually asks, “am I contagious now,” that behaves like lowered specificity. Two tests, one virus, each trustworthy in the direction the other is not.

The old clinical mnemonic says it compactly: a highly sensitive test’s negative rules out; a highly specific test’s positive rules in. Which test you are holding decides which direction of the reversal you can trust. The fault is not in the tests. It is in reading either one’s advertised number as an answer to your question.

Dinnes J et al. Rapid, point-of-care antigen tests for diagnosis of SARS-CoV-2 infection. Cochrane Database Syst Rev 2022, CD013705: symptomatic sensitivity 73.0% (95% CI 69.3 to 76.4), specificity 99.1%. On PCR positivity outlasting infectiousness: Cevik M et al. Lancet Microbe 2021;2:e13–e22.

Why it happens

Tests are advertised by the direction that is measured, because that is what a laboratory can validate: take known infected and known healthy samples, run the test, count. The patient’s direction cannot be measured that way, because it depends on who is being tested. A test with fixed accuracy gives a different answer to “what does my positive mean” in a hospital ward than at a family dinner, because the condition is common in one place and rare in the other. So the honest number is the one that gets printed, and the reader supplies the reversal without noticing that a second number was required.

What to ask

- Which direction does that accuracy figure run: the chance of the result given the condition, or the chance of the condition given the result?
- How common is this condition in people like me? Without that, the reversal cannot be done.
- Was this test built to catch cases or to clear people? That tells you which of its results to trust.
- For a screening result: out of a thousand people like me who got this result, how many turned out to have it?