

How can a positive test be wrong?

Juan Segurola · All Rights Reserved

QUE · Article 047 · Simple Questions, Deep Answers · Version 0.1 EN

A positive result can occur when the condition being tested for is absent. Even a test that correctly classifies a large fraction of cases can generate many such results when the condition is uncommon in the tested population. To understand a positive result, we need to count both the true positives and the false positives that could have produced it.

The following is an invented numerical case, not a clinical test or advice about anyone's diagnosis. It uses natural frequencies so that the relevant groups remain visible. Research on reasoning has examined how this form of representation can make conditional-inference tasks easier, but the arithmetic here stands independently of a claim about how well any particular reader will perform. [1]

1. Start with the population, not the result

Imagine 10,000 items being checked for a specified defect. An independent reference procedure establishes that 100 are defective and 9,900 are not. The prevalence of the defect is therefore 1% in this population. Assume for this exercise that the reference classification is correct.

Now suppose the screening test has sensitivity 90%: it is positive in 90 of the 100 defective items. It misses the other 10. Suppose its specificity is 95%: it is negative in 9,405 of the 9,900 nondefective items. The remaining 495 nondefective items give positive results.

We have not claimed that 90% or 95% of positive results are right. Those percentages have different denominators. Sensitivity starts among the defective items; specificity starts among the nondefective items. The question raised by a positive result starts among the items that tested positive.

2. Change the denominator

There are $90 + 495 = 585$ positive results. Only 90 correspond to the defect. The positive predictive value in this hypothetical population is therefore $90/585$, approximately 15.38%.

The 495 false positives are not evidence of fraud or of a test that does nothing useful. They arise from the assumed 5% false-positive rate acting on a very large nondefective group. The 90 true positives come from a high detection rate acting on a much smaller defective group. Both counts must enter the calculation.

This is the conceptual trap: “the test detects 90% of defects” answers a question about known defective items. It does not answer “given a positive result, what fraction of items are defective?” Reversing the conditioning changes the question, even though both sentences contain the same words, positive and defective.

The calculation is exact for the stipulated counts. In an actual study, sensitivity, specificity and prevalence would be estimated, with uncertainty and possible bias. The neat integers here should not be mistaken for that degree of empirical certainty.

3. Change one assumption at a time

Keep the test's sensitivity and specificity fixed, but let prevalence rise to 10%. Among 10,000 items, 1,000 are now defective and 9,000 are not. We expect 900 true positives and 450 false positives, so positive predictive value becomes $900/1,350$, or about 66.67%.

Nothing about the assumed test has improved. What changed was the composition of the tested population. A result interpreted in one population cannot simply inherit the predictive value measured in another. Holding test characteristics constant is itself a modelling assumption; in practice those characteristics can also differ across populations and operating conditions.

Return to prevalence 1%, but now increase specificity to 99%, keeping sensitivity 90%. There are still 90 true positives, but now only 99 false positives. Positive predictive value becomes $90/189$, about 47.62%. This substantial improvement still does not make a positive result certain.

These two variations separate two levers: changing whom we test and changing the false-positive rate. They are not interchangeable interventions, and neither calculation tells us what operating threshold or testing policy is appropriate for a real application.

4. Why repeating the test is not automatically a solution

Suppose a second test has the same rates and, conditional on the true state, its errors are independent of the first test's errors. At prevalence 1%, the expected number of defective items positive on both is $100 \times 0.9 \times 0.9 = 81$. The expected number of nondefective items positive on both is $9,900 \times 0.05 \times 0.05 = 24.75$. The corresponding fraction is $81/105.75$, about 76.60%.

The non-integer 24.75 is an expected count across repetitions of the hypothetical sampling process, not a fractional physical item. More importantly, independence is doing crucial work. If the same interference causes both tests to be positive, multiplying the error rates can greatly overstate the benefit of repetition.

At the opposite extreme, if the second test simply reproduces every first-test result, two positives contain no more information than one. The calculation remains $90/585$. Real follow-up procedures require evidence about dependence, sample handling and the reference standard; repetition alone does not supply that evidence.

5. What remains outside this answer

This example treats a binary state and a binary result. It does not decide how to define a real condition, where to set a threshold on a continuous measurement, or how to weigh missed cases against false alarms. Those decisions require additional evidence and explicit consequences.

A false positive here is also not a p-value. It is an incorrect positive classification relative to the stipulated reference state. Confusing that event with a research hypothesis test, or with deliberate fabrication, would change the subject.

The practical intellectual lesson is to ask for the relevant population and denominators before interpreting the result. A positive result is evidence produced by a process with specified error rates, not a guarantee. In this worked case, the answer becomes clear when we count the alternative routes by which an item can enter the positive group.

Related reading

Article 046, “Why can adding a road make traffic worse?”, also shows why an apparently favourable local fact does not determine the final outcome without the surrounding conditions.

References

[1] Hoffrage, U., Krauss, S., Martignon, L., and Gigerenzer, G. “Natural frequencies improve Bayesian reasoning in simple and complex inference tasks.” *Frontiers in Psychology* 6, 1473 (2015).
<https://doi.org/10.3389/fpsyg.2015.01473>