Harvard Data Science Review • Special Issue 1: COVID-19: Unprecedented Challenges and Chances

Building Intuition Regarding the Statistical Behavior of Mass Medical Testing Programs

Lance Waller[1] Taal Levi[2]

[1]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia, United States of America,

[2]Quantitative Wildlife Ecology, Conservation, and Environmental Genetics Lab, Department of Fisheries Wildlife and Conservation Sciences, College of Agricultural Science, Oregon State University, Corvallis, Oregon, United States of America

The MIT Press

Published on: May 14, 2021

DOI: https://doi.org/10.1162/99608f92.19de8159

ABSTRACT

The quality of medical decision-making and public health planning alike depends directly upon understanding the accuracy of medical tests, especially during a pandemic. But the statistical concepts and measures used to assess test accuracy can be confusing. Why is there not one single definitive measure of test accuracy? How much should individuals worry about spreading COVID-19 if their test results are negative? What do sensitivity, specificity, false positive results, false negative results, and positive predictive value mean relative to each other? In this tutorial, we clarify the meaning of these terms in intuitive ways via visual illustrations, and explain how these terms are all connected to one another through Bayes' theorem. We show how to use the relationships in that theorem to assess personal risk when large numbers of people are being tested. We illustrate as well the extent to which the accuracy of large numbers of tests depends on the proportion of those tested who have the disease. Overall, we aim to heighten a general intuition regarding the performance of mass medical testing campaigns. Here, toward that end, we review different ways to measure the accuracy of diagnostic tests with reference to pandemic-specific examples.

Keywords: medical testing, diagnostic testing, sensitivity, specificity, positive predictive value, false positives

---

1. Introduction: The Vocabulary of Outbreaks

The COVID-19 pandemic caused a massive spike in the broad-scale, government-implemented diagnostic testing of large numbers of people. Researchers and officials have used the data generated by such testing programs to highlight two important numbers: namely, those showing the *incidence* and *prevalence* of the disease. The *incidence* describes how many new cases occurred in a recent time period, for example, the past week. The *prevalence* describes how many people are currently infected right now. Prevalence may be larger than incidence as it includes new cases from the past week but also individuals who were infected during the prior week and have not yet recovered. Important though these two numbers are, however, they are only part of a larger story that can be difficult for researchers, health workers, and the general public to fully understand.

The planning of broad-scale testing programs is challenging to implement, and the results can be challenging to interpret. To begin with, the task of making sense of test reports at the individual, regional, national, and international levels requires connecting different types of data, different types of tests, and different assessments of test performance. Beyond that, the terminology used in this context can present obstacles. For example, press reports detailing the results of broad-scale testing programs often rely on words such as 'sensitivity' and 'specificity,' which connote 'accuracy' in everyday usage. But in epidemiologic reports that describe the performance of testing programs, sensitivity and specificity have different, specialized meanings. Hence, sorting through statements regarding test performance in varied media can be difficult, especially when

seemingly counterintuitive results are quoted without much explanation. Two recent examples vividly demonstrate this challenge:

> "You might think any test with 95 percent sensitivity and 95 percent specificity would be highly accurate. But while these would be great grades on an organic chemistry final, the ability of such a test to render a reliable result is extremely poor: 50 percent of the positive results would not be true positives, Dr. Osterholm said. (You'll have to take my word for this—explaining the statistics would require half a column!)" (Brody, 2020)

> "The CDC explains why testing can be wrong so often. A lot has to do with how common the virus is in the population being tested. 'For example, in a population where the prevalence is 5%, a test with 90% sensitivity and 95% specificity will yield a positive predictive value of 49%. In other words, less than half of those testing positive will truly have antibodies,' the CDC said.…Alternatively, the same test in a population with an antibody prevalence exceeding 52% will yield a positive predictive value greater than 95%, meaning that less than one in 20 people testing positive will have a false positive test result." (CNN Wire, 2020.)

Both press reports make true but confusing statements, and implicitly prompt questions. How can a test with two measures of performance at or above 90% be wrong more than half the time? Constraints on space and scope in popular media rarely allow writers and readers to develop insight or build intuition into why these seemingly counterintuitive associations occur. It is no wonder that nonexperts may end up being confused. Indeed, even statisticians and epidemiologists familiar with terms such as sensitivity and specificity often struggle to explain them to colleagues and friends.

Here, we aim to provide definitions and examples to illustrate the relationships between epidemiologic terms such as sensitivity, specificity, and prevalence. These relationships are important for the design and assessment of testing strategies, programs, and protocols. Building intuition around how and why certain strategies work well (or not) is critical for monitoring the ongoing pandemic, communicating results on current status, and targeting vaccine rollout to areas of high incidence and prevalence. Misunderstanding or miscommunicating testing results can result in inefficiencies at best and increased mortality at worst.

In considering the examples below, we suggest keeping two key questions in mind regarding any reported percentage: 'What question does that percentage answer?' and 'From what population is the percentage taken?' We consider first the impact of decisions regarding who is tested on the numbers of cases reported. Next, we consider the main qualities of an 'accurate' test (that is, sensitivity and specificity), and explain how and why these qualities are related.

2. Patterns in Data Reflect Patterns of Disease and Patterns of Testing

Through daily reports on the number of individuals testing positive for COVID-19, *public health surveillance* provides critical information regarding which absolute counts of infection, hospitalization, and death are increasing, which are leveling off, and which are declining. Surveillance data typically come from a variety of sources, including testing, hospital records, and mortality records. The regional and temporal patterns observed in combined surveillance data pertain not only to infection but also to varied modes of reporting and testing, such as what types of data are included, which individuals are tested, how quickly regions and hospitals report data, and whether the separate data elements cover similar time periods.

When an individual is tested for an infectious disease, the test result typically determines whether a person has or does not have evidence of infection. Some diagnostic tests report whether the individual is currently infected, and others test for antibodies or other clues that the individual was infected in the past. When we add up results for individuals at the regional or national levels, it is important to consider the patterns of testing and of the testing results. Early in the pandemic, when resources were scarce, testing focused on individuals with symptoms, those with known contacts with infected individuals, and frontline health care workers.

Patterns of tested individuals within a given city/region/state/country reflected local decisions and priorities for testing: specifically, which tests were most important for making critical decisions at that point in time in that region? For example, when local tests were limited *and* the policy goal was to identify the number of severe cases that then currently required or would soon require hospitalization, the testing of asymptomatic individuals may have been viewed as less of a priority than the rapid identification of severe cases and contact tracing.

In this scenario of limited testing, the (appropriate) focus of tests on those with severe symptoms can result in *selection bias* with respect to simple estimation of the proportion of population currently infected via the proportion of positive tests among those tested. As an example, suppose a clinic has 10 remaining test kits and 20 people waiting to be tested, seven of whom currently have a high fever and are short of breath (symptoms suggestive of COVID-19). For simplicity, suppose the other 13 patients are not infected. In order to prioritize treatment for the sickest individuals, we would assign the first seven remaining COVID-19 tests to these seven symptomatic individuals and then test three of the remaining individuals. If seven of the 10 tests were positive for COVID-19, the percent testing positive would be 70% (seven positives out of 10 tests); however, only seven out of 20 people in line (35%, seven infected out of 20 people) were actually infected. In this example, the individuals with severe symptoms were (again, appropriately at the time) more likely to be tested and more likely to be positive, yielding a proportion of positive tests much higher than the proportion of infected individuals. Here, as elsewhere, accuracy in understanding and reporting the proportions of positive tests would have required reporting the number of positives out of the number tested *and* reporting the priorities by which individuals were chosen to be tested.

As tests became more available over the course of the pandemic, policy goals shifted to the assessment of what proportion of the regional population was infected at a given time (the current *prevalence* of the disease, including both new and ongoing infections). Generally speaking, to avoid selection bias, the most accurate estimates of current prevalence should involve testing a random sample of the population containing both infected individuals and uninfected individuals. The proportion of sampled individuals who are infected would then accurately reflect the proportion of infected individuals in the population (the prevalence). Very few efforts were put in place to develop national- or state-level sampling-based estimations of infection prevalence. And, eventually, as tests became more widely available, the goal shifted from testing random individuals in a resource-limited environment to testing *everyone* in particular settings when testing had become more widely available (e.g., on university campuses, the state of Georgia, the U.S. military). Some public health researchers extended these ideas and proposed even broader proposals for mass application of rapid tests.

At the national or international level, these changing goals led to a global map of information influenced by different patterns of testing in different areas. In order to facilitate accurate interpretation of changing local prevalence values over time, analysts, decision makers, and the interested public need to understand these shifting priorities for testing and data collection over the course of the pandemic.

3. What Are the Qualities of an 'Accurate' Test?

Diagnostic tests are not perfect. Some uninfected people will have positive test results. Such results are referred to as 'false positive results' or 'false positives.' In addition, some infected people will have negative test results, referred to as 'false negatives.' In this context, we can observe one of four outcomes. Two of these are correct outcomes:

- A true positive test result: an infected individual has a positive test.
- A true negative test result: an uninfected individual has a negative test.

The other two possibilities reflect different types of incorrect outcomes:

- A false negative test result: an infected individual has a negative test.
- A false positive test result: an uninfected individual has a positive test.

An 'accurate' test will have a high proportion of correct outcomes and low proportions of each type of incorrect outcome. But in practice, a low proportion of false positives does not necessarily mean a low proportion of false negatives. As an extreme example, suppose we have a test that *always* reports a positive result, regardless of whether the tested individual is infected or not. Such a test would have no false negative results, but it would be a useless and, indeed, harmful test. The same would be true of a second test that *always* reports a negative result. These two examples suggest the need for some sort of balance between the risks of different types of incorrect outcomes.

Different measurements of the performance of a diagnostic test focus on different aspects of test performance such as the proportion of false positive or false negative results; no single measure captures the full set of performance characteristics we desire in a good diagnostic test. The examples of the always-positive or always-negative tests considered here provide examples of tests with wonderful (perfect!) performance on one measure, and terrible performance on the other.

Specific definitions of test performance based on the proportion of false positives or the proportion of false negatives appear in many biostatistics and epidemiology textbooks. These are often used as examples introducing Bayes' theorem, a key mathematical formula from probability theory that can be used to define the relationships between the different measures of testing accuracy (McGrayne, 2018). Bayes theorem is presented as a 'gotcha' homework problem in introductory biostatistics and epidemiology classes to show students how some very common intuitions about probability are actually very wrong. Here, rather than present the theorem first as an equation to explain associations between different measures of test accuracy, we will instead begin by illustrating the associations through examples and then summarizing these with the theorem.

Probability abounds with confusing examples and seeming 'paradoxes,' which is why gambling is a profitable venture for the house. Many of these paradoxes result in a difference between what we *want* to happen (for example, 'I think this slot machine is due for a payout'), and what the randomization mechanism driving outcomes actually generates (each spin is equally likely to pay out). The setting of mass testing is no different. We *want* a test with good properties in one arena to be good in another. As a result, we often attempt to attribute good results in one subset of performance measures to the rest.

4. Illustrations With COVID-19 Testing

To better understand different measures of test performance and their interrelationships, we consider the example of two broad types of testing for COVID-19. The first type of test is used to identify an active infection by extracting RNA, converting it into complementary DNA, and assessing whether that DNA (but no other) can be amplified using a technique called real-time polymerase chain reaction. A second type of tests seeks biological markers of immunity within an individual due to prior exposure rather than an active infection by assessing the presence of antibodies to the virus. The latter test (often referred to as the 'antibody test') led some federal and state government officials to propose that mass antibody testing can identify those who have been infected and are presumably immune and safe to return to work (Mukherjee, 2020). (See *CDC Guidance on Antibody Testing*,). In short, the RNA test determines if the individual is *currently infected* and the antibody test determines if an individual has been *infected in the past*. Additional rapid testing approaches continue to be under active development. For ease of description in the sections below, we focus on the antibody test, although the concepts apply to tests of current infection as well.

The first antibody test approved for emergency use by the Food and Drug Administration (FDA) was made by Cellex. Cellex's test reports a *sensitivity* of 93.8% and a *specificity* of 95.6%. We use these values to define and

interpret these two summaries of test performance.

In the definitions below, we consider several different proportions of different subpopulations. *The key to keeping the definitions straight is tracking which subpopulations go with which measures of performance*. We note that this is easier said than done.

To begin, the *sensitivity* of the antibody test is the probability that a person *tests positive* (T_+) given they have COVID-19 antibodies, which can be written as a conditional probability: $Pr[T_+|D_+]$ (where we read "$T_+|D_+$" as an individual tests positive (T_+) given that they indeed are "disease positive" (D_+), that is, they have antibodies). For the antibody test, *sensitivity* answers the question: *What proportion of individuals test positive among the subpopulation of individuals with antibodies? In other words, how often is the test correct for people who have antibodies?*

A sensitivity of 93.8% sounds very good since it means that only 6.2% (100% – 93.8%) of people who actually had COVID-19 receive a (false) negative test result. This 6.2% represents the proportion *of tested individuals with antibodies* who would falsely presume that they are still susceptible and take unnecessary care not to become infected (e.g., stay out of work when in fact they could safely return if long-term immunity holds).

Next, *specificity* is the probability that a person tests negative (T_-) given that they are 'disease negative' (D_-), denoted as $Pr[T_-|D_-]$. *Specificity* answers the question: *What proportion of people test negative among the subpopulation of healthy people? In other words, how often is the test correct for people who don't have antibodies?*

A *specificity* of 95.6% indicates that only 4.4% (100.0% – 95.6%) of 'healthy' people (who don't have protective antibodies and are still susceptible) would receive a (false) positive test. In the setting of the COVID-19 antibody test, this seeks to address the question, 'Have I been infected in the past?' In this case, 4.4% of the healthy population would falsely assume that they have immunity due to prior infection with the virus when, in fact, they are still susceptible.

What do sensitivity and specificity tell us about how one should interpret a positive test result? Since the sensitivity is high, most people receiving a positive antibody test result would assume that they are positive and immune to COVID-19. Recall that sensitivity measures the proportion of people *testing* positive in the subpopulation of people who *are* positive (have antibodies), that is, $Pr[T_+|D_+]$. However, a person who just received a positive test result wants to know the proportion of people who *are* indeed positive *among the subpopulation of people with positive tests*, that is, they want to know $Pr[D_+|T_+]$.

In this example, *what you know* (my test is positive) and *what you wish to know* (do I have antibodies?) are reversed from the definition of sensitivity given above. In probability terminology, we say that the conditioning is reversed, and as such answers a different question. The probability you are positive given that you test

positive can be written as $\Pr[D_+|T_+]$ and is called the *positive predictive value*, which reverses the conditioning from sensitivity, $\Pr[T_+|D_+]$.

Similarly, the *negative predictive value*, the proportion of individuals with negative tests who are, in fact, negative can be written as $\Pr[D_-|T_-]$. This reverses the conditioning from the specificity, $\Pr[T_-|D_-]$, *the* proportion of individuals who are in fact negative who receive negative test results.

The conditional relationships of which proportion is taken from which subpopulation is particularly confusing for two other terms used to describe the performance of mass diagnostic testing. The *false discovery rate*, denoted as $\Pr[D_-|T_+]$, defines the proportion of people with positive tests who are negative. This reverses the conditioning from the false positive proportion, denoted $\Pr[T_+|D_-]$, defined by the proportion of people who are negative but receive positive test results. In sum, the false discovery rate describes the proportion of 'discoveries' (positive tests) that are false (occur in healthy people), while the *false positive proportion* describes the proportion of healthy people who receive (false) positive test results. The first is a proportion of positive tests; the second is a proportion of healthy people.

4.1. Illustration 1: Grouping by Antibodies or Grouping by Test Results?

As noted above, keeping track of which term refers to which proportion of which subpopulation is a challenge, even for experienced statisticians and epidemiologists! To help clarify the situation, consider Figure 1 (building on similar figures in McKenna, 2020). Here, figures with red heads truly have the disease (D_+) and figures with blue heads are healthy (D_-). The squares held by each figure represent a piece of paper with that individual's test result. Red squares denote a positive test result (T_+) and blue squares denote a negative test result (T_-). Note that there are four possibilities and that all combinations can occur (positive people can receive positive or negative tests and negative people can receive positive or negative tests). To keep things straightforward in this hypothetical example, suppose we use a test with 80% sensitivity and 90% specificity (actual COVID-19 tests typically have higher values). Supposing we test 10 people with the disease (red heads) and 10 people without the disease (blue heads), we see that *sensitivity* represents a proportion among people with the disease and specificity represents a proportion among healthy people. That is, *sensitivity* and *specificity* are defined by a proportion among individuals grouped by their true disease status. In contrast, if we rearrange these same 20 people's tests (10 with the disease and 10 healthy) by the test results they received (nine total positive tests and 11 total negative tests) we see that the positive predictive value, the false discovery rate, and the proportion of false negatives are all defined as proportions among individuals with the same test outcome.

Figure 1 clarifies that the terms used to define test performance are based on different subpopulations of interest. As noted above, sensitivity defines a proportion of individuals among the subpopulation who 'are positive', that is, they truly have antibodies. Specificity defines a proportion of individuals among the subpopulation who 'are negative,' that is, truly don't have antibodies. While both terms are reported as

percentages, they are percentages of different subpopulations. To further clarify these relationships, we summarize these terms, their definitions, and probability notation in Table 1.

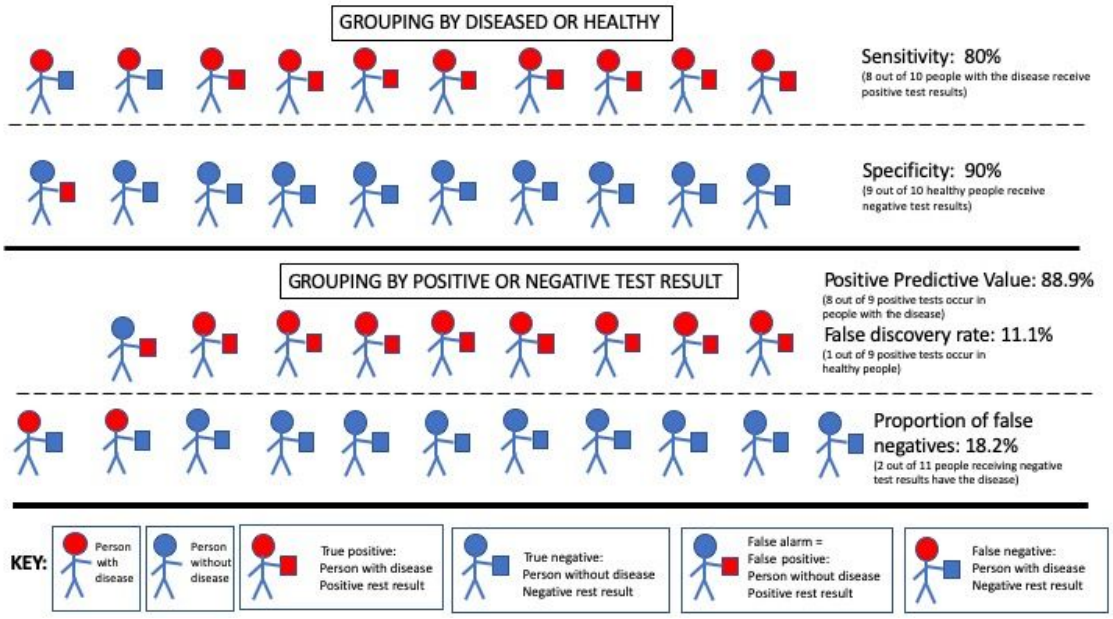

Figure 1. Definitions of test outcomes when grouping individuals by health status (sensitivity and specificity) or by test result (positive predictive value, false discovery rate, and proportion of false negatives). (See text for additional discussion). *Note. Figure updated May 17, 2021, to correct typographic error.*

Table 1. A list of terms referring to specific aspects of test performance.

Term	Probability notation	Definition
Sensitivity	$\Pr[T_+ \mid D_+]$	The proportion of people who test positive *among the subpopulation who are positive*
False negative proportion	$\Pr[T_- \mid D_+]$	The proportion of people who test negative *among the subpopulation who are positive.*
Note: Since *sensitivity* and the *false negative proportion* are proportions of the same population and since individuals either test positive or test negative, *sensitivity + false negative proportion* = 100%		
Specificity	$\Pr[T_- \mid D_-]$	The proportion of people who test negative *among the subpopulation who are negative*

False positive proportion	$\Pr[T_+ \mid D_-]$	The proportion of people who test positive *among the subpopulation who are negative.*
Note: Since *specificity* and the *false positive proportion* are proportions of the same population and since individuals either test positive or test negative, *specificity + false positive proportion* = 100%		
Positive Predictive Value	$\Pr[D_+ \mid T_+]$	The proportion of people who are positive *among the subpopulation who test positive.*
False discovery rate	$\Pr[D_- \mid T_+]$	The proportion of people who are negative *among the subpopulation who test positive.*
Note: Since *positive predictive value* and the *false discovery rate* are proportions of the same population and since individuals either are positive or are negative, *positive predictive value + false discovery rate* = 100%		
Negative Predictive Value	$\Pr[D_- \mid T_-]$	The proportion of people who are negative *among the subpopulation who test negative.*
Prevalence	$\Pr[D_+]$	The proportion who are positive *across the entire population tested.*

Note. Here, T_+ and T_- *i*ndicate that an individual receives a positive or negative test, respectively, while D_+ and D_- *indicate that an individual does or does not have the disease the test is trying to detect. The notation* $\Pr[A \mid B]$ *indicates the conditional probability of A being true given B is true. We can interpret this to mean the proportion of individuals where A is true among the subpopulation of individuals where B is true.*

Table 1 also includes the *prevalence* of a disease, the proportion of individuals currently having the disease. In assessing the performance of the PCR (population currently infected) test for COVID-19, the prevalence refers to the proportion of individuals currently infected among those tested (including new *and* existing cases), while for the antibody test, the prevalence refers to the total proportion of individuals who currently have antibodies. In a mass testing situation, the prevalence is defined to be the proportion *among those tested*. This may be different from the proportion of the entire population with antibodies due to testing priorities and strategies. This clarification is necessary since, as discussed above, COVID-19 tests are allocated based on testing priorities, due to local resources, availability, and access.

The prevalence of a disease in the tested population affects the associations between sensitivity, specificity, and positive predictive value. While we typically think of sensitivity and specificity as properties of a particular

test, in a large-scale mass testing setting their relationships to each other change with the background prevalence of disease. These relationships change because *prevalence changes the relative sizes of the subpopulations defining each measure of test performance*.

To illustrate this connection, note that in Figure 1 we earlier assumed a prevalence of 50% in those tested (we had 10 red heads with antibodies and 10 blue heads without antibodies). Now, consider Figure 2 where we apply the same tests (80% sensitivity and 90% specificity) to a prevalence of 20% in the tested population (10 red heads with antibodies and 50 blue heads without antibodies). As before, we first arrange people by disease status (people with antibodies together and people without antibodies together), then we have *the same people* rearrange themselves by test result (positive tests together and negative test together).

When we rearrange by test result, we find that the positive predictive value has changed from 88.9% in Figure 1 to 61.5% (only 8 out of 13 positive tests occur in people with antibodies). As we test more people without antibodies, the number of false positive test results begins to increase. The *percentage* of positive tests in individuals without antibodies remains the same, but since we are testing many more people without antibodies, the *number* of such false positive test results increases. If we were to test even more people without antibodies (say, 100, 1,000, or 1,000,000) but still tested 10 people with antibodies (that is if the prevalence of people with antibodies decreases in the tested population), at some point we could (and would) observe more false positives (positive test results in people without antibodies) than we would observe true positives (positive test results in people with antibodies)!

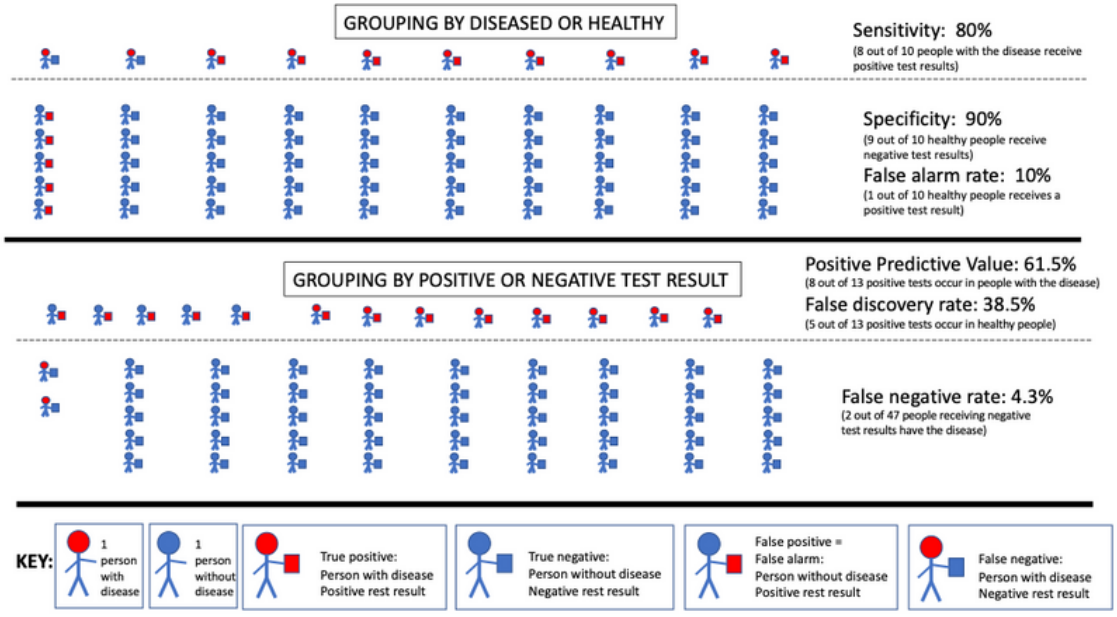

Figure 2. Definitions of test outcomes in a lower prevalence setting than Figure 1. Individuals remain grouped by health status (sensitivity and specificity) or by test result (positive predictive value, false discovery rate, and false negative rate). (See text for additional discussion).

Figures 1 and 2 suggest that there is a specific relationship linking sensitivity, specificity, positive predictive value, and prevalence of the disease within the tested population. The figures suggest that the key to understanding the relationship among these quantities depends on how we group people into subpopulations, the relative sizes of these subpopulations, and the proportions of incorrect tests within each subpopulation. We next use three separate but related descriptions ranging from conceptual to mathematical to illustrate these relationships in more detail. Each example provides a bit more insight into the interrelationships and their relevance to better design, implementation, and understanding of the results for large-scale diagnostic testing.

4.2. Illustration 2: Tracking Possible Outcomes

As a next step, we move away from the simplified settings of Figures 1 and 2, and consider examples with prevalence, sensitivity, and specificity closer to examples arising from COVID-19.

Our goal is to track all of the possible outcomes of testing that could happen and arrange these in a 'tree' or flowchart of possible events (Figure 3). The branches on the flow chart provide more detail on subgrouping than we showed in Figures 1 and 2.

To better match local testing for COVID-19, consider a population of 10,000 people tested with a true *prevalence* (proportion of individuals who are antibody positive among those tested) of only 1%, similar to values early in the pandemic. Here, this means that 1% of the population tested previously had COVID-19 and has antibodies. The prevalence can be expressed probabilistically as $\Pr[D_+] = 0.01$, in which case $\Pr[D_-] = 0.99$. These add to 100% since each individual either does (D_+) or does not have antibodies (D_-). Of the original 10,000 people, the prevalence means 100 individuals are truly positive and so 9,900 are truly negative. An antibody testing sensitivity of 93.8% means about 94 of these 100 D_+ individuals will (properly) receive a positive test result. (Recall that sensitivity is a proportion of those who truly have the antibodies.) The specificity of 95.6% means that 9,464 of the 9,900 negative individuals will (properly) receive negative test results. Only 4.4% of truly negative individuals will receive false alarms and test positive, but this corresponds to 4.4% out of a large number (9,900) of negative people. The result is 436 people are falsely told that they have antibodies to the disease and are presumably immune (but in fact are not). The proportion of positive tests that are true positives is then the 94 true positive tests divided by the total number of positive test results (true and false). As noted above, this quantity, $\Pr[D_+ | T_+]$, gives the *positive predictive value* and equals 0.18 (Figure 1C). Similarly, the *false discovery rate*, or $\Pr[D_- | T_+]$, is 0.82. Thus, at the 1% prevalence level, only 18% *of people who test positive* will actually be positive, and 82% of people who test positive and think they have antibodies will be wrong!

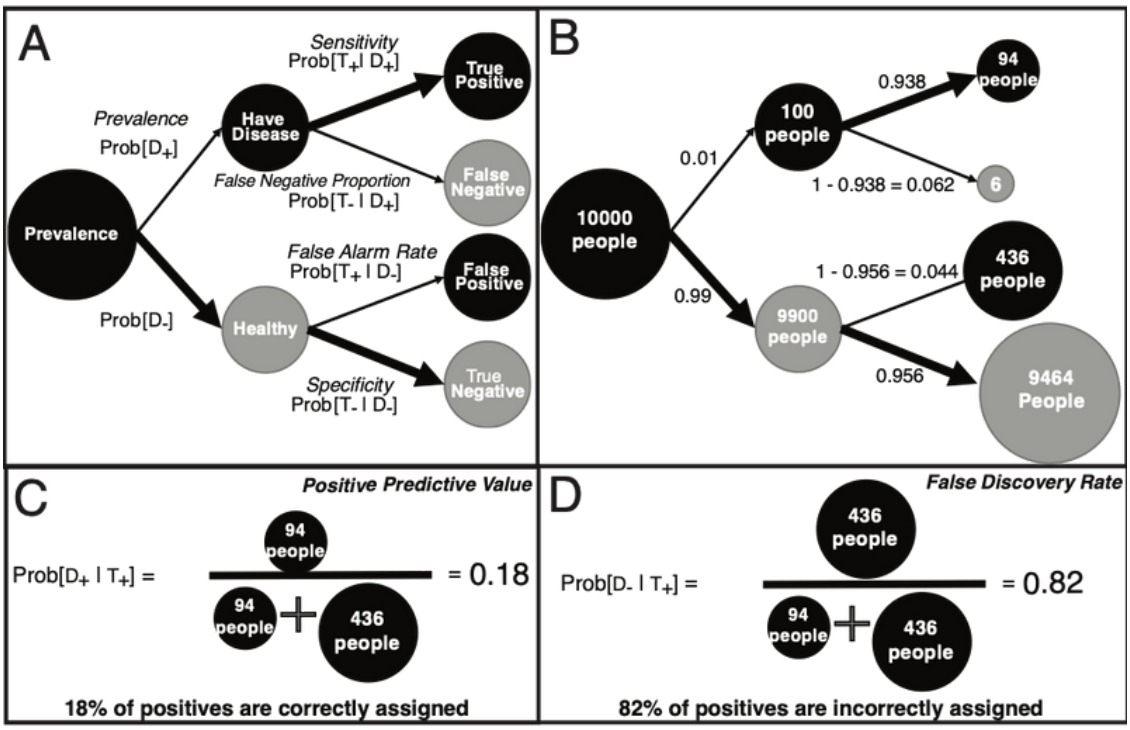

Figure 3. A tree of possible events. (A) Events delineated based on probability notation and epidemiological terms. (B) The expected distribution of 10,000 people through this tree of events when prevalence is 1%, sensitivity is 93.8%, and specificity is 95.6% (based on specifications of the first antibody test to achieve Food and Drug Administration [FDA] emergency use authorization). The probabilities and numbers can be used to calculate (C) the positive predictive value, and (D) the false discovery rate.

While Figure 3 tracks all of the possible outcomes, the results above could be summarized into a single sentence such as: 'A test with sensitivity of 93.8% and specificity of 95.6%, when applied to a low-prevalence population, can yield a positive predictive value of only 18%!' As we mentioned in the introduction, similar sentences have frequently appeared in press coverage of COVID-19 testing. Based on our discussions so far, we can start to see how to explain the truth hidden in such a statement. The key is to note that the sentence lists three percentages representing proportions of three different subpopulations (those with antibodies, those without antibodies, and those testing positive, respectively). This key is apparent only if the definitions of the proportions defining sensitivity, specificity, and positive predictive values are known and readily at hand for all readers.

The example in Figure 3 demonstrates that *most* people who receive a positive test result can be negative if the prevalence is low in the tested population. This reinforces the suggestion from Figures 1 and 2 that the impact of the background *prevalence*, $\Pr[D_+]$, is essential to understanding the performance of a mass testing campaign.

Why would this matter? For some diseases, a false positive result simply results in a healthy person receiving an unnecessary treatment. If the treatment is not too toxic, then this is not particularly consequential. However, in the case of COVID-19, where prevalence among the tested was small early in the pandemic, the association between the different test criteria can be critical. If the prevalence of the disease among those tested is very small, we see that the majority of people who test positive for antibodies will falsely think that they are immune. Under this mistaken assumption, they may leave themselves open to the risk of severe disease and death, as well as maintaining transmission among the general public.

In order to provide insight on how decisions regarding whom to test can influence the performance of a mass testing strategy, we recall that the prevalence represents the proportion *of individuals tested* who have antibodies. If individuals are selected from the general public in a random sample (such that every person is equally likely to be tested), then the prevalence in the tested population will be a close approximation to the overall prevalence in the antibodies in the general population. If, however, we focus testing on individuals who have had close contact with infected individuals, the *prevalence in the tested population* will be higher than the *prevalence in the general population* and our positive predictive value will improve. While we may not know the precise prevalence in the general population, we may exercise some control over the prevalence in the tested population by focusing on tests for higher risk individuals.

Next, we examine how sensitivity and specificity affect the relationship between prevalence among those tested and the positive predictive value of a test.

4.3. Illustration 3: How Prevalence Changes Positive Predictive Value

To clarify the impact of prevalence on the positive predictive value, we extend the COVID-19 example used in Figure 3 for prevalence ranging from 0% to 100%. In Figure 4, we plot the positive predictive value (the probability that an individual has antibodies given their test was positive) against the prevalence among tested individuals. The dark line represents the relationship for *sensitivity* of 93.8% and a *specificity* of 95.6% (the values from our COVID-19 examples above). There is a sharp decrease in the positive predictive value as the prevalence among the tested population drops (Figure 4A). Next, we redraw this relationship with a new curve for increasing values of sensitivity. We see that the positive predictive value is almost completely unchanged as we increase sensitivity from 93.8% to 95%, 96%, 97%, 98%, and 99% (overlapping red lines in Figure 4B). The blue lines in Figure 4C, however, show that changes in specificity from 95.6% to 96%, 97%, 98%, to 99% lead to substantially higher positive predictive values at any given prevalence value.

This example demonstrates that we can maintain a better positive predictive value for the tested population if: (1) we focus a testing program on those at higher risk (that is, we aim to have a higher prevalence of antibodies among the tested individuals than we would if we tested at random) and (2) if we increase test *specificity* (that is, if we have a choice between different tests, we choose the one with higher specificity). The plots also show that the positive predictive value is relatively unaffected by changes in test *sensitivity*.

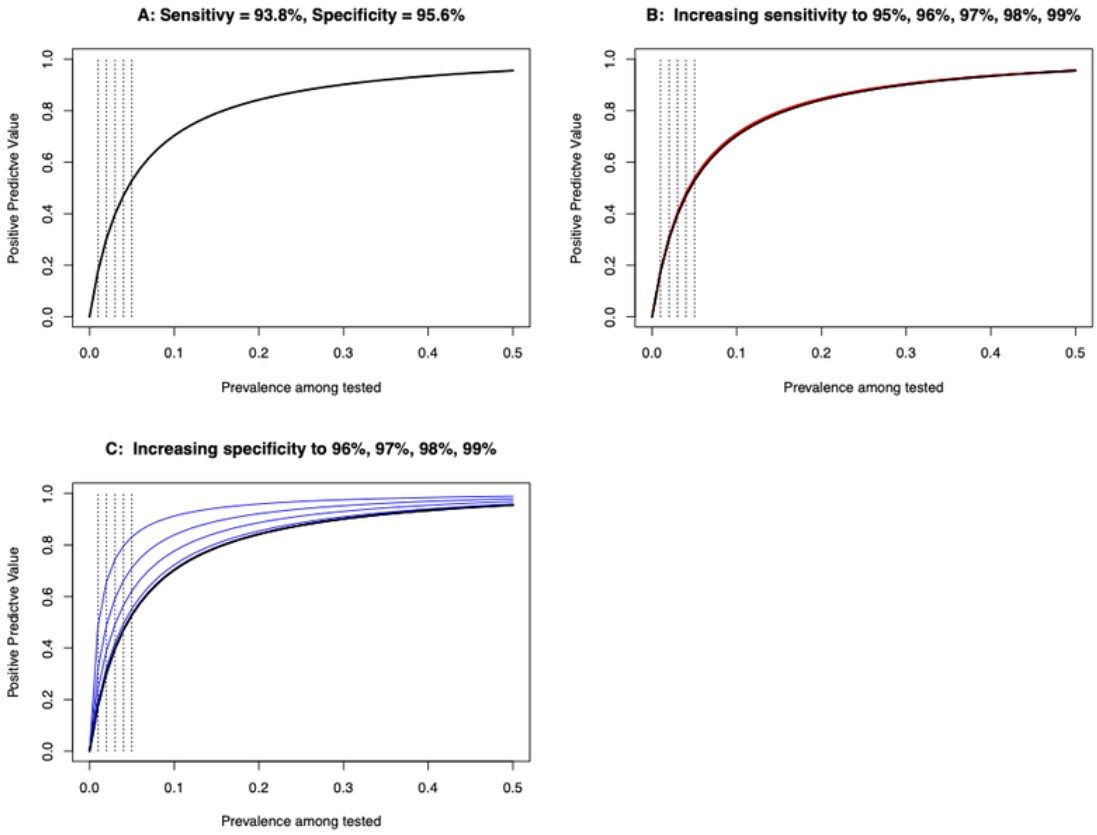

Figure 4. The relationship between the prevalence of antibodies among those tested and the positive predictive value. In (A), we see increasing positive predictive value with increasing prevalence (prevalence of 1% to 5% indicated by vertical dashed lines. In (B), we see the minimal impact of increasing sensitivity from 95% to 99% (red lines). In (C), we see the improvement in positive predictive value associated with increasing specificity from 96% to 99% (blue lines).

The curves in Figure 4 further stress the critical role of prevalence in assessing the performance of a mass testing campaign. Ongoing, sampling-based prevalence estimates can (and should) provide improved situational awareness of the proportion of the population currently infected (for PCR tests) or infected sometime in the past (antibody tests). However, assuming the absence of such surveillance-based estimates of the current level of COVID-19 prevalence at the national or local levels, Figure 4 still provides important information for planning and expectations of performance for population-level testing because it illustrates the impact of changes in sensitivity and specificity for any value of prevalence. For example, if the best information suggests a prevalence (current or past) of 5 to 10%, we can examine the relationships in Figure 4 across this range to gain insight on the performance of the testing program.

5. The Formulas Behind the Figures

As mentioned above, the probability tool defining the relationships between our testing performance measures is Bayes' theorem, best understood as a consequence of the mathematical definition of conditional probability.

First notice that the joint probability that someone both is positive and tests positive $\Pr[\mathrm{D}_+ \text{ AND } \mathrm{T}_+]$ can be written as the probability of being positive given a positive test $\Pr[\mathrm{D}_+ | \mathrm{T}_+]$ multiplied by the probability of receiving a positive test independent of your antibody status $\Pr[\mathrm{T}_+]$. This relationship can be written as

$$\Pr[\mathrm{D}_+ \text{ AND } \mathrm{T}_+] = \Pr[\mathrm{D}_+ | \mathrm{T}_+] \Pr[\mathrm{T}_+].$$

Similarly, the joint probability of being positive and testing positive can be written as,

$$\Pr[\mathrm{T}_+ \text{ AND } \mathrm{D}_+] = \Pr[\mathrm{T}_+ | \mathrm{D}_+] \Pr[\mathrm{D}_+].$$

Setting these expressions equal to each other (since $\Pr[\mathrm{T}_+ \text{ AND } \mathrm{D}_+] = \Pr[\mathrm{D}_+ \text{ AND } \mathrm{T}_+]$) and using a little algebra, we find

$$\Pr[\mathrm{D}_+ | \mathrm{T}_+] = \frac{\Pr[\mathrm{T}_+ | \mathrm{D}_+] \Pr[\mathrm{D}_+]}{\Pr[\mathrm{T}_+]}.$$

This formula can be expressed using the epidemiological terms for test performance as follows: the *positive predictive value (*$\Pr[\mathrm{D}_+ | \mathrm{T}_+]$) is proportional to the *sensitivity (*$\Pr[\mathrm{T}_+ | \mathrm{D}_+]$) multiplied by the *prevalence* of antibodies among those tested ($\Pr[\mathrm{D}_+]$). The denominator, $\Pr[\mathrm{T}_+]$, is the probability that a randomly selected person receives a positive test, which can be calculated as the probability of a positive test for the people who test positive and are positive plus the probability of a positive test for people who test positive and aren't. In the notation above, this can be written as,

$$\Pr[\mathrm{T}_+] = \Pr[\mathrm{T}_+ | \mathrm{D}_+] \Pr[\mathrm{D}_+] + \Pr[\mathrm{T}_+ | \mathrm{D}_-] \Pr[\mathrm{D}_-].$$

Taken together, we have

$$\text{Positive Predictive Value} = \frac{\text{Sensitivity} \times \text{Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + \text{False Positive Proportion} \times (1 - \text{Prevalence})}$$

where False Positive Proportion is 1 – Specificity (Table 1). Using the Cellex antibody test values for sensitivity (93.8%) and specificity (95.6%) and assuming a prevalence of 1%, we can place the appropriate values into Bayes' theorem to obtain

$$\Pr[\mathrm{D}_+ | \mathrm{T}_+] = \frac{0.938 \times 0.01}{0.938 \times 0.01 + 0.044 \times 0.99} = 0.18,$$

which is the same result we calculated in Illustration 21 by following the tree of possible outcomes (Figure 3C). Thus, whether by tracing possible outcomes or by considering conditional probability and implementing Bayes rule, 82% of people who test positive for COVID-19 would be false positives (1 – 0.18) if the prevalence among those tests was 1%. By changing prevalence values, Bayes' theorem provides the values to graph the curves in Figures 4A, 4B, and 4C in Illustration 3.

6. Discussion

The descriptions above illustrate concepts, graphs, and formulas summarizing the challenges of different but related measures of diagnostic performance in a large-scale, mass diagnostic testing program. Given these challenges, we must take care when reading, writing, and understanding descriptions of testing systems to frame issues in terms of the specific questions answered and to consider how sensitivity and specificity relate to the expected proportions of results in different subpopulations. Claims, reports, and manuscripts that deal with testing systems must be considered in the context of the prevalence of the relevant outcome of interest, for example, current levels of infection (PCR tests) or antibodies (antibody tests) within the tested population.

For example, when we read statements such as ‘the test is 99% accurate’ we should immediately think of two questions: (1) How does the author define ‘accurate’? and (2) ‘99%’ of which subpopulation? If ‘99% accurate’ refers to individuals with positive disease status (antibodies or active infection), then ‘accurate’ refers to sensitivity. If ‘99% accurate’ instead is for individuals without negative disease status, then ‘accurate’ refers to specificity. We have seen above that these are two important but different qualities that lead to two different interpretations of the statement.

The connections between sensitivity, specificity, positive predictive value, false discovery rate, and prevalence are essential for understanding the performance of testing strategies. The illustrations here show that if the prevalence of the disease among those tested is small, the small proportions of false results from tests with seemingly high specificity can result in low positive predictive values. Taken cumulatively, the examples show that this occurs because a small *percentage* of false positive results can result in a large *number* of false positive results if we are testing a large number of individuals without the outcome of interest (low prevalence). This issue is not something we can remove with adjusted calculations but is, in a sense, ‘baked in’ to mass testing settings.

The impact of prevalence on mass testing programs sometimes is referred to as the ‘base-rate fallacy’ (where the base-rate corresponds to the prevalence among those tested) and has a long history in the diagnostic testing literature (Bar-Hillel, 1980). When designing a mass testing system, we can protect against this problem to some degree by choosing tests with higher specificity or by opting to test a greater proportion of those likely to have the disease (for example, testing those with confirmed contacts with infected individuals) thereby yielding a higher prevalence among tested individuals (Service, 2020; Watson & Whiting, 2020; Woloshin et al., 2020). Other adjustments in practice include multistage testing wherein all positive tests are followed up by a second round of testing. By focusing the second round of testing on only those with positive results in the first round, we are effectively testing a subgroup with higher prevalence, with the effect that each round can improve in performance. Such approaches have been implemented in many university settings where positive antigen tests are followed up with ‘more accurate’ PCR tests that have higher specificity.

While our examples are based on COVID-19, the same issues arise whenever large groups of people are tested. For example, consider the question of whether we should extend recommendations for routine mammogram screening for breast cancer for women or PSA screening for prostate cancer in men to younger ages? Like most cancers, the prevalence of both of these types of cancer increases with age. Adding younger individuals to the testing pool will lower the overall prevalence in the tested population and, as seen above, will lower the positive predictive value of tests. These are challenging decisions: Each case detected early is important, but many false positives will likely sour the population on participating in screening programs.

The definitions and descriptions here provide tools for exploring what levels of prevalence in the testing population will provide adequate performance for both individual tests and the testing strategy in general. In summary, the evaluation of testing programs requires clear thinking and careful reporting. While a statement such as 'Applying a test with 99% sensitivity and 95% specificity can result in over 50% false positives!' makes for provocative reading, it compares different percentages of different subpopulations (individuals with antibodies, individuals without antibodies, and those receiving positive tests) and it omits any mention of prevalence. When writing, we should be very careful to specify definitions and the groups to which percentages refer. When reading, we should ask 'of what?' after every percentage, and recall the specific questions that each term answers as well as the specific subpopulation to which each term refers.

Disclosure Statement

Lance Waller and Taal Levi have no financial or non-financial disclosures to share for this article.

References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Brody, J. E. (2020, July 13). The case for smarter coronavirus testing. *The New York Times*. https://www.nytimes.com/2020/07/13/well/live/coronavirus-smart-testing.html

CNN Wire. (2020, May 27). Antibody tests for COVID-19 wrong up to half the time, CDC says. *Fox40 News*. https://fox40.com/news/coronavirus/antibody-tests-for-covid-19-wrong-up-to-half-the-time-cdc-says/

McGrayne, S. B. (2011). *The theory that would not die*: *How Bayes' rule cracked the Enigma code, hunted down Russian Submarines, and emerged triumphant from two centuries of controversy*. Yale University Press.

McKenna, S. (2020, May 5). What COVID-19 antibody tests can and can't tell us. *Scientific American.* https://www.scientificamerican.com/article/what-covid-19-antibody-tests-can-and-cannot-tell-us/

Mukherjee, S. (2020, May 27). The CDC says antibody tests shouldn't be used to make return-to-work decisions. *Fortune*. https://fortune.com/2020/05/27/antibody-test-coronavirus-covid-19-cdc-return-to-work-testing-immunity/

Service, R. F. (2020). Fast, cheap tests could enable safer reopening. *Science, 369*(6504), 608–609. https://doi.org/10.1126/science.369.6504.608

Watson, G. P., & Whiting, P. F. (2020). Interpreting a Covid-19 test result. *British Medical Journal*, *369*(8245), Article m1808. https://doi.org/10.1136/bmj.m1808

Woloshin, S., Patel, N., & Kellelheim, A. S. (2020). False negative tests for SARS-CoV-2 infection—Challenges and implications. *New England Journal of Medicine, 383*, Article e38. https://doi.org/10.1056/NEJMp2015897
