Diagnostic Testing for COVID-19
Considering False Positive and False Negative Results

Deborah A. Zarin, MD[1]
Joseph Lau, MD[2]

[1] Multi-Regional Clinical Trials Center
Brigham and Women's Hospital
And Harvard University
Boston, Massachusetts, USA

[2] Center for Evidence Synthesis in Health
Brown University School of Public Health
Providence, Rhode Island, USA

Background

The COVID-19 global pandemic has brought attention to diagnostic tests. The lack of accurate, available and timely testing for COVID-19 has been blamed for the size and severity of the pandemic within the United States. And now that social distancing and related public health actions have demonstrated that they can "flatten the curve", many are focusing on how diagnostic tests can help us emerge from the current lockdown conditions. [1]

The sheer volume of medical information disseminated each day through peer-reviewed literature, pre-print servers, government agencies, news reports, and even Twitter, is unprecedented and overwhelming. New information, of variable quality, appears daily in all these venues. We focus here on tests designed to detect the virus, and tests designed to detect host antibodies to the virus. These tests are the foundation of many current policies and proposals for relaxing lockdown conditions[2]. A casual reader might feel reassured by reports of sensitivity and specificity values that sound "high" (e.g., in the 90s). Our goal here is to help readers understand how to critically evaluate such reports against the backdrop of the current pandemic, and especially how to think about proposed clinical or public health policies that rely on these tests.

At this point in the pandemic, many laboratories offer tests to detect viral RNA or to detect antibodies; levels of standardization, validation and regulatory oversight are highly variable. We aim to elucidate key principles of diagnostic test interpretation by using prototypical data for commonly used diagnostic tests for COVID-19. Many issues need to be considered when evaluating studies of diagnostic tests, including but not limited to: proper selection of the reference standard, spectrum effects, and verification bias.[1] The urgency of applying available tests in the rapidly developing pandemic has circumvented some of these important principles. As a result, many aspects of the tests remain poorly understood. Nonetheless, these tests have to be used to help manage patients and control the pandemic despite knowledge gaps. Here, we focus in particular on concerns about inaccurate results when tests are used to make decisions about individuals. Given the rapidly evolving nature of tests and their performance, we do not intend to provide up-to-date data on test characteristics, and front-line workers should not rely on these data for their current decision making. In addition, the use of these tests for epidemiologic surveillance is critical, but is not discussed here.

To remind readers of the basic principles of test performance, the tables use prototypical values of sensitivity and specificity for the two categories of diagnostic tests (detecting virus and detecting antibodies) , along with a range of prevalence values, to calculate the expected numbers of false positives and false negatives. The Figure shows the impact of prevalence on the positive predictive value of positive and negative test results.

Points to consider when reading about diagnostic tests:

1. The potential value of a test depends on how one interprets a "true positive" and a "true negative." At this early stage in the pandemic, scientists do not yet know the

exact relationship between what the tests are designed to measure, the clinical situation of interest, and how these relationships can change over the course of the illness. For example, how does the presence of virus in the nasopharynx at various points in the illness relate to infectiousness? Similarly, how does the presence and perhaps quantity of antibodies in serum relate to clinical immunity, and does this vary over time? In other words, even if our current tests were 100% accurate at detecting what they are designed to detect, we would still not be sure who was infectious and who was immune.

2. In the ideal test, all "positive" results would mean that the patient has the condition being tested for, and all "negative" results would mean that the patient does not have the condition being tested for. In reality, tests are not perfect; they all produce some false results. Three factors influence the frequency of false results: prevalence of the condition in the population being tested; the sensitivity of the test; and the specificity of the test. The Figure and the tables show how these factors interact.

3. The utility of a test can only be assessed in the context of its specific intended use. For example, different settings will likely involve different prevalence estimates for the condition of interest (e.g., the general, asymptomatic population vs. hospitalized patients who are considered highly likely to have COVID-19). In addition, the impact of "false positives" and "false negatives" will depend on what judgments or decisions are being made based on the test results. A test that produces many false negatives may be very concerning in some situations, and much less concerning in other situations.

Common Scenarios:

1. <u>Determining which patients require COVID-level of infection control protocols:</u>
RT-PCR tests based on nasopharyngeal swabs are being used, in part, to guide decisions about infection control protocols in healthcare settings. One possible use of this test is to determine if healthcare personnel treating a person under investigation require full personal protection equipment; similarly, one might use such a test to determine when a person who previously tested positive might be able to be moved to an environment with a lower level of infection control. In both scenarios, the high specificity of the test means that there would be very few false positive results. However, the biggest concern in this scenario is false negatives, since these could lead to inadequate protection for staff who are unknowingly treating a patient who is still infectious. The probability that patients have COVID-19 infection is high in this scenario. The sensitivity of nasal swab-based PCR tests might be limited by the sampling method, and the degree to which COVID-19 patients have virus in the nasopharynx at various points in their illness. Current estimates put the sensitivity in the 65-75% range, and specificity of about 95%[3]. Assuming a PCR test sensitivity of 75%, Table 1 illustrates the impact of prevalence. For example, if the prevalence were presumed to be 90%, then the test would produce 225 false negatives for every 95 true negatives; in this setting, there is still a 70% chance of being infectious even after a negative test. A lower prevalence

would produce dramatically fewer false negatives. It can be seen that the "value" of the test depends on how it is used, and how good the clinical assessment of "risk of COVID" is in the intended circumstance.

2. Determining who can return to work

   PCR tests are being discussed as a method of screening workers prior to allowing them to enter the workplace. If asymptomatic workers (who have not had known direct exposure to a COVID-19 patient) are screened, then the presumed prevalence would be relatively low. In this case, false negatives will be less common than in more highly prevalent settings of a hospital, but could nonetheless pose a risk of exposing others in the workplace, the public with whom they interact, and ultimately with their family and other close contacts. Given the proposed wide-spread use of such a testing strategy, even a small rate of false negatives could lead to a large number of workers who are falsely reassured that they are free of the virus.

   Serology tests for antibody status are also being proposed as a mechanism for identifying people who could safely return to work, perhaps in public-facing positions. These tests have been reported to have both sensitivity and specificity in the range of 95% (see Table 2)[4]. If such tests were used for the general population of workers who have no reason to believe that they have had COVID-19, then the presumed prevalence is likely to be low. For a population with a prevalence of 10%, there would be approximately 50 false positives for every 10 true positives. As with the false negative PCR tests, the false positive serology tests pose the concern that falsely reassured people will be put into situations in which they could become sick and also could then infect others. However, the use of this test in the situation described is likely to produce very few false negatives (fewer than 1 per 1000 people tested); this is helpful given the desire to enable as many people as possible to safely return to work and the community.

What to watch for in the daily deluge of information?

First, it is important to watch for emerging information about the relationships between "true" test results and the clinical conditions of interest (infectiousness and immunity). Second, remember that tests are coming online faster than they can be evaluated. Consumers of medical information have to keep an eye open for information about test performance (sensitivity and specificity) from reputable sources. Third, each proposed use must be thought through: what would a true positive or negative mean? What would be the impact of a false positive or negative? How likely are false results, given the prevalence and the known test characteristics? Given the economic and social burdens of the social distancing policies, it is understandable that much hope is being placed on the use of diagnostic tests. However, hope cannot replace scientific understanding and empirical data. There are no perfect tests, and we must use any test with our eyes open to the likelihood and impacts of false results.

References:

1. Smetana GW, Umscheid CA, Chang S, Matchar DB. Editorial: Methods Guide for Authors of Systematic Reviews of Medical Tests: A Collaboration Between the Agency for Healthcare Research and Quality (AHRQ) and the Journal of General Internal Medicine. In: Chang SM, Matchar DB, Smetana GW, Umscheid CA, eds. *Methods Guide for Medical Test Reviews*. AHRQ Methods for Effective Health Care. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012. http://www.ncbi.nlm.nih.gov/books/NBK98247/. Accessed April 20, 2020.

2. National coronavirus response: A road map to reopening. *American Enterprise Institute - AEI*. https://www.aei.org/research-products/report/national-coronavirus-response-a-road-map-to-reopening/. Accessed April 21, 2020.

3. Wang W, Xu Y, Gao R, et al. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA*. March 2020. doi:10.1001/jama.2020.3786

4. administrator J website. Global Progress on COVID-19 Serology-Based Testing. Johns Hopkins Center for Health Security. https://www.centerforhealthsecurity.org/resources/COVID-19/serology/Serology-based-tests-for-COVID-19.html. Accessed April 17, 2020.

Table 1. Frequency of true and false results using different prevalence estimates, assuming the Covid-19 PCR test sensitivity is 75% and test specificity is 95%.

Prevalence	Number of people out of 1000 with			
	true+[1]	false+[2]	true-[3]	false-[4]
1/1000 (0.1%)	0.75	49.95	949.05	0.25
10/1000 (1%)	7.5	49.5	940.5	2.5
100/1000 (10%)	75	45	855	25
500/1000 (50%)	375	25	475	125
900/1000 (90%)	675	5	95	225

[1] True positive = prevalence*sensitivity
[2] False positive = (1 – prevalence)*(1 – specificity)
[3] True negative = (1 – prevalence)*specificity
[4] False negative = prevalence)*(1 – sensitivity)

Table 2. Frequency of true and false results using different prevalence estimates, assuming the serology test sensitivity is 95% and test specificity is 95%.

Prevalence	Number of people with out of 1000			
	true+	false+	true-	false-
1/1000 (0.1%)	0.95	49.95	949.05	0.05
10/1000 (1%)	9.5	49.5	940.5	0.5
100/1000 (10%)	95	45	855	5
500/1000 (50%)	475	25	475	25
900/1000 (90%)	855	5	95	45

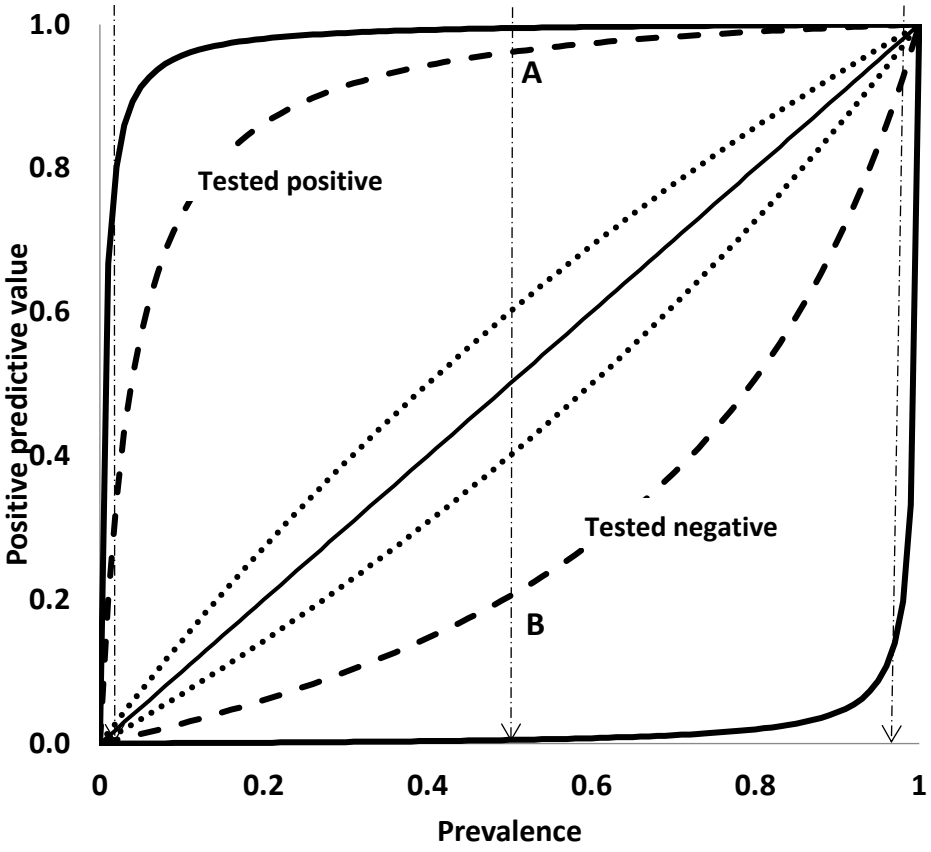

Figure. This graph illustrates the impact of disease prevalence on the positive predictive value of a test. The diagonal line depicts a useless test that provides no discriminatory value; in other words, the probability that a person has the condition being tested does not changed based on the test results. A set of two curves moving away from the diagonal provides incremental discriminatory information as the sensitivity and specificity of a test increase. The dot pair of lines based on a test with sensitivity of 60% and specificity of 60% depicts the positive predictive value of a test that returns a positive result (upper line that convex toward left upper corner). This is a test with poor performance characteristics that is unlikely ever used in clinical practice. The positive predictive value of a test that returned a negative result is shown as the corresponding lower curve that points toward lower right corner. The middle pair of curves (dashed lines) depicts a test with sensitivity of 75% and specificity of 97% (test performance similar to that of current COVID-19 PCR test). The outer pair of curves (solid lines) depicts a test with very high sensitivity of 99.5% and very high specificity of 99.5% (close to being a perfect test). The sensitivity and specificity values are used for illustrative purpose only. Three vertical lines (dot-dash) are drawn (at 2%, 50%, and 98%) to help interpret test result at a specific prevalence. For example, points A and B represent the positive predictive value of a positive test result and a negative test result, respectively for the test with 75% sensitivity and 97% specificity. A positive test result in a patient with an estimate chance of 50% of having COVID-19 will bring the after test probability greater than 90%. A negative test result will reduce the chance of infection to less than 20%. This test offers the greatest discriminatory ability in this range of prevalence; whereas little value is gained at the extreme prevalence.