

Special Issue on the COVID-19 Pandemic with Guest Editors Oren Caspi, M.D. and Ami Neuberger, M.D.

The Sky Has Its Limits in COVID-19 Testing

Shai Linn, M.D., Dr.P.H.^{1*†}, Shay Tzafrir, Ph.D.^{2*}, and Shay Gueron, Ph.D.^{3*}

¹School of Public Health, The University of Haifa, Haifa, Israel; ²School of Management, The University of Haifa, Haifa, Israel; and ³Department of Mathematics, The University of Haifa, Haifa, Israel

ABSTRACT

At the time of writing, in July 2020, the COVID-19 pandemic has already inflicted dramatic international restrictions, including airports closing and limiting international travel. It has been suggested that re-opening of airports should involve and even rely on testing travelers for COVID-19. This paper discusses the methodology of estimating the detection and diagnostic accuracy of COVID-19 tests. It explains the clear distinction between the technical characteristics of the tests, the detection measures, and the diagnostic measures that have clinical and public health implications. It demonstrates the importance of the prevalence of COVID-19 in terms of determining the ability of a test to yield a diagnosis. We explain the methodology of evaluating diagnostic tests, using the predictive summary index (PSI), and the minimum number of tests that need to be performed in order to correctly diagnose one person, which is estimated by $1/PSI$. In a population with low prevalence, even a high-sensitivity test may lead to a high percentage of false positive diagnoses, resulting in the need for multiple high-cost tests to achieve a

Abbreviations: COVID-19, coronavirus disease 2019; *fnf*, false negative fraction (in artificial population); *FNF*, false negative fraction (in real population); *fpf*, false positive fraction (in artificial population); *FPP*, false positive fraction (in real population); *J*, Youden index; *NND*, number needed to diagnose; *nns*, number needed to screen; *NNT*, number needed to treat; *NPV*, negative predictive value; *PPV*, positive predictive value; *PSI*, predictive summary index (Ψ); *RD*, risk difference; RT-PCR, reverse transcriptase–polymerase chain reaction; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Citation: Linn S, Tzafrir S, Gueron S. The Sky Has Its Limits in COVID-19 Testing. Rambam Maimonides Med J 2020;11(3):e0020. doi:10.5041/RMMJ.10412

Copyright: © 2020 Linn et al. This is an open-access article. All its content, *except where otherwise noted*, is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Acknowledgement: This research was partly supported by the Israel Science Foundation (grant No. 3380/19) and the Center for Cyber Law and Policy at the University of Haifa, in conjunction with the Israel National Cyber Directorate in the Prime Minister's Office.

Conflict of interest: No potential conflict of interest relevant to this article was reported.

* All authors contributed equally to the writing of the manuscript.

† To whom correspondence should be addressed. E-mail: linnshai@gmail.com

correct diagnosis. Thus, basing a policy for opening airports on diagnostic testing, even with the best test for COVID-19, has some limits.

KEY WORDS: Bayes' theorem, COVID-19, diagnostic tests, epidemiology, flights, methods, screening

INTRODUCTION: WHY DO WE NEED TO STUDY TEST ACCURACY?

At the time of writing, in July 2020, more than 13 million persons worldwide have been infected by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), leading to a pandemic of coronavirus disease 2019 (COVID-19). More than half a million persons have died of the disease.¹ The accurate detection of the viral infection by medical and laboratory tests is critical for finding a cure and for planning public response measures. However, the results that the tests provide depend on the prevalence of the disease, as is explained below.

All medical tests are inaccurate, i.e. they may yield both false positive and false negative results. Some of these errors may have grave consequences for individuals and society. A false negative diagnosis may lead a person to be falsely assured and therefore avoid necessary quarantine. On the other hand, a false positive diagnosis may lead to unnecessary quarantine or unnecessary treatment.

We discuss the properties of the associated calculations, the accuracy measures of medical tests, and the estimate of the minimum number of tests that need to be performed in order to correctly diagnose one patient.

TWO TYPES OF TESTS FOR COVID-19

There are two main types of tests used for detecting and diagnosing COVID-19 infection for clinical or surveillance purposes: molecular and serological.²⁻⁴

Molecular tests detect the pathogen of acute disease. These tests may also detect fragments of the pathogen before it is fully cleared from the body, even if the pathogen is no longer able to replicate or cause disease. Currently, such tests rely on a technique called reverse transcriptase–polymerase chain reaction (RT-PCR) to detect the presence of the virus. These tests can provide data on the incidence of the disease, i.e. the fraction of a population that is newly infected. This paper is focused on the evaluation of these tests.

Serological tests—sometimes referred to as “antibody tests”—can provide information about the prior infection of an individual, indicating the body’s immune response to the virus, and can detect the infection after convalescence. These tests can provide data on the prevalence of the disease, i.e. the fraction of the population that has been infected by the disease in the past. The body makes IgM and then IgG antibodies within about 10 days; IgM indicates a very recent infection, and IgG indicates infection at any time in the past. Serological tests can help public health officials determine the prevalence of previous infection, including among asymptomatic individuals, and among those with mild symptoms who did not seek medical care. Prevalence determination may help in deciding to relax social distancing and quarantine measures. Further, the serological test may be used to explore an individual’s previous infection by the virus. This paper explains the importance of utilizing the information from serological tests, i.e. for determining the prevalence of the disease.

TWO TYPES OF TEST ACCURACY MEASURES

Each of the above tests may be used for two different goals: *detection* (Table 1) or *diagnosis* (Table 2) of the disease.⁵⁻²¹

Detectability Measures: Technical Characteristics of No Clinical Importance

We use italic lower-case letters in the description of screening in the general population in a 2×2 table, Table 1. The *sensitivity* and *specificity* are calculated in samples of persons with (a+c) and without (b+d) the disease in a selected population. In this table, it is not appropriate to include the totals of the “horizontal” axis of test (T) results.

A researcher could determine the detectability of COVID-19 in a study where the prevalence of the disease is *artificial*. For example, a study may calculate the *sensitivity* and *specificity* of a test in 100 persons with a disease (e.g. clinical COVID-19), and

Table 1. Data Presentation in a Selected Population, Assessing the Detection Capability of a Test.

		COVID-19	
		S+ Sick	S- Not Sick
Test Results	T+ Positive Test	a	b
	T- Negative Test	c	d
		a+c	b+d

$$sensitivity = P(T+ | S+) = \frac{a}{a+c}, \quad specificity = P(T- | S-) = \frac{d}{b+d}$$

Error terms for the study population:

$$false\ positive\ fraction = f_{pf} = \frac{b}{b+d}$$

$$false\ negative\ fraction = f_{nf} = \frac{c}{a+c}$$

$$Youden\ Index = J = sensitivity + specificity - 1$$

nns = minimal number of people needed to be screened to detect correctly one person = $1/J$

100 persons without the disease. We set the prevalence of COVID-19 in this particular study, artificially, to be 50%.

The *sensitivity* and *specificity* are used to describe the *technical* characteristics of a test. These measures are not useful in the clinical setting, because the prevalence of the disease is different in a true patient population. The sensitivity or specificity could tell us the *percentage* of the persons with (or without) the disease that would be detected, but we will not know *how many* patients with COVID-19 (or without it) will be diagnosed correctly. For example, we could know the percentage of persons with the disease that would be quarantined based on detection by a test, but we could not know how

many persons would be quarantined. The information on the percentage of detected persons would be meaningful only if we know the prevalence of the disease.

The fraction (percent) of persons with the disease who would not be identified (i.e. the false negative fraction) is $f_{nf} = 1 - sensitivity$. The fraction (percent) of persons without the disease who would not be identified (i.e. the false positive fraction) is $f_{pf} = 1 - specificity$.

The Youden index (*J*) is a summary measure of the goodness of a test. It describes the percent of correct detection (without false negative nor false positive detection). This index is defined as:

Table 2. Data Presentation in a Clinical Study Setting in a Target Patient Population Assessing the Diagnostic Capability of a Test.

		COVID-19		
		S+ Sick	S- Not Sick	
Test Results	T+ Positive Test	A	B	A+B
	T- Negative Test	C	D	C+D

$$PPV = P(S+ | T+) = \frac{A}{A+B}, \quad NPV = P(S- | T-) = \frac{D}{C+D}$$

Error terms for the study population:

$$\text{False Positive Fraction} = FPF = \frac{B}{A+B} = 1 - PPV$$

$$\text{False Negative Fraction} = FNF = \frac{C}{C+D} = 1 - NPV$$

$$PSI = \Psi = PPV + NPV - 1$$

$$NND = \text{Minimal Number Needed to be tested to diagnose correctly one person} = 1/PSI$$

$$\begin{aligned} J &= 1 - (fpf + fnf) \\ &= 1 - [(1 - \text{sensitivity}) + (1 - \text{specificity})] \\ &= \text{sensitivity} + \text{specificity} - 1 \end{aligned} \quad (\text{Eq. 1})$$

When $J=1$, the test is always correct: there are no errors, so $fpf + fnf = 0$, and the test detects correctly the sickness status.

When $J=0$, assuming that *sensitivity* and *specificity* are of equal importance in determining the expected gain, the test provides no information. In other words, the test is useless if the proportion of errors equals 100%, and $fpf + fnf = 1$, leading to $J=0$.

When $-1 < J < 0$, the test is misleading: its results are negatively associated with a true diagnosis. When $J = -1$, the test is always misleading.

J can also be interpreted as the difference between the true and false positive fractions.

Since $J = \text{sensitivity} - fpf$, J reflects the *excess* of the proportion of a *positive* result among patients with the disease versus patients without the disease. Similarly, J also reflects the excess in the proportion of a *negative* result among patients without the disease versus patients with the disease. This can also be written as $J = \text{specificity} - fnf$.

J as a difference measure of detectability analogies to a cohort study

Table 1 is analogous to a clinical trial or a cohort study that compares the risk of a disease among those exposed to a risk factor, R_{exposed} , and the risk among those who are not exposed, $R_{\text{non-exposed}}$. The “causative” variable (i.e. the “exposure”) is the fact that a person does or does not have the disease, and the diagnostic test results (positive or negative) are the “outcome” of the disease. The difference in risk between the exposed and non-exposed persons is measured by the risk difference (RD):

$$RD = \text{Risk difference} \\ = R_{\text{exposed}} - R_{\text{non-exposed}} \quad (\text{Eq. 2})$$

J is analogous to RD .

$$J = \text{sensitivity} - \text{fpf} \quad (\text{Eq. 3})$$

Therefore, a derived analogy of the well-known measure of the “number needed to treat”, (NNT) = $1/RD$, is $1/J$. The value $1/J$ may be interpreted as the number of persons that need to be examined in order to correctly detect by screening (nns) one person with the disease (Table 1) of persons with and without the known disease. The nns could help in estimating the minimum number of tests that has to be applied to persons with *known* diagnosis of COVID-19 (with or without the disease) in order to detect one person correctly (positive or negative, respectively). It can be useful in assessing a *percent* of a successful monitoring program (how many of the persons with, or without, the disease will be detected). However, it cannot assess *how many* persons with or without the disease will be detected, and thus it has no clinical or public health importance, because it cannot be applied to a real population in which we do not know the COVID-19 diagnoses.

Currently, PCR tests have a sensitivity or specificity of approximately 70%–95%, depending on the conditions of the tests.⁵ For example, the sensitivity of the PCR test using a nasopharyngeal swab is higher than that using a nasal swab, while the specificity of the test is lower using a nasopharyngeal swab. For simplicity, we assume here a 90% sensitivity and 90% specificity for both PCR and serological tests.

Example

Suppose that a population of 1000 travelers is screened for COVID-19, using the PCR test. The test will detect 90% of the persons with COVID-19. These persons will be treated or quarantined. However, the test will not detect 10% of the persons with the disease, i.e. the test will have an *fnf* of 10%. This would allow 10% of the persons with COVID-19 to continue interacting with their family and the community, with the implied risk of transmitting the disease. Similarly, the test would correctly detect 90% of the persons without COVID-19. These persons would not be quarantined. However the test would incorrectly detect infection in 10% of the uninfected persons, i.e. the test will have an *fpf* of 10%. This would allow 10% of the persons without COVID-19 to be unjustifiably quarantined. For these data, J would be 0.80. This indicates 80% better information compared to the case where no test is used, or if a useless test with $J=0$ is used. Note that we know only the test’s ability to detect the percentage, and not the number of persons that would be diagnosed with or without the disease. We know that the test will detect 90% of the persons with (or without) the disease. However, since the prevalence of the disease in the traveler population is not known, we will not know the number of persons with the disease (or without the disease, respectively). The test would not show how many travelers would need to be quarantined, nor how many travelers would not be quarantined because of false negative test results. Based solely on detection measures, it is not possible to plan quarantine facilities or to assess the number of persons with COVID-19 who are wrongly released into the community and continue to infect others. Thus, the detection measures cannot be useful for practical planning of public health measures. By contrast, diagnostic measures, which are explained below, can be used for these purposes.

Diagnostic Measures of Clinical and Public Health Importance

The application of a diagnostic test to a patient (target) population utilizes a 2×2 table (Table 2). To evaluate the effectiveness of the application of a diagnostic test in the patient population, the investigator first observes the outcome, i.e. the test results, and obtains information about the study factor, i.e. the disease status.

We use *upper-case letters* to describe screening in the patient population in a 2×2 table, Table 2. It is the data in this table that are of interest to the patient (and the physician) and public health officials, answering the following questions: (1) When the test is positive, what is the probability that the patient has the disease? (answerable by the positive predictive value [PPV]); (2) When the test is negative, what is the probability that the patient does not have the disease? (answerable by the negative predictive value [NPV]).

In this clinical setting, the diagnoses are as yet unknown, and the test is used to diagnose COVID-19 in individuals: the PPV and NPV are an estimate of the test's ability to diagnose patients accurately in a population (based on the real disease prevalence), i.e. of the fractions of patients who are *diagnosed correctly* as positive or negative, respectively. The PPV is the fraction (percent) of the positive tests in a given population that will correctly diagnose a COVID-19 patient. Similarly, NPV is the fraction (percent) of negative tests that will correctly diagnose a person who is not infected. The fraction (percent) of persons with a positive test who would not have the disease is the diagnostic false positive fraction (FPF), that is calculable as $FPF = 1 - PPV$. The fraction (percent) with a negative test result who have the disease and is diagnosed incorrectly as not having the disease is the diagnostic false negative fraction (FNF) calculable as $FNF = 1 - NPV$.

Both PPV and NPV depend on the proportion of the population that has the disease according to clinical or serological criteria, i.e. the *prevalence* of the disease. Thus, the PPV and NPV provide insight into the expected accuracy of the positive and the negative test results in a given population, by factoring in the ability of the test to detect the disease and the prevalence of the disease in the population.

Suppose that the same test were used in two different populations: population A with a higher disease prevalence and population B with a lower disease prevalence. Then, the PPV would be higher in population A than in population B, because the number of false positives would be a lower percentage of the total number of positive tests in population A. Similarly, the NPV would be higher in population A than in population B.

Predictive summary index as a summary measure of diagnostic ability of a test in individuals

A summary index, the predictive summary index (PSI, ψ), is a measure of the additional information given by the test results, beyond the prior knowledge (the prevalence of the disease).²¹ Note that the information from a positive test result beyond what is already known about the disease prevalence is $PPV - Prevalence$. Similarly, the information from a negative test result beyond what is already known about the probability of no disease (the prevalence of no disease) is $NPV - (1 - Prevalence)$.

Thus, the overall information, i.e. the gain in certainty obtained after a test is performed, beyond what is already known, can be calculated as a summary measure:

$$\begin{aligned} \text{Total gain in certainty} &= \\ &= [PPV - Prevalence] + \\ &+ [NPV - (1 - Prevalence)] = \\ &= PPV + NPV - 1 = PSI \end{aligned} \quad (\text{Eq. 4})$$

Alternatively, ψ is a summary of the information that is not derived from errors, FNF and FPF:

$$\begin{aligned} PSI &= 1 - (FPF + FNF) = \\ &= 1 - [(1 - PPV) + (1 - NPV)] = \\ &= PPV + NPV - 1 \end{aligned} \quad (\text{Eq. 5})$$

If $\psi=1$, the test is always correct: there are no errors, so that $FPF + FNF = 0$; i.e. the test detects correctly the sickness status. When $PSI = 0$, $PPV + NPV = 1$, and the test provides no overall information. In other words, the test is useless if the proportion of errors equals 100%; i.e. when $FPF + FNF = 1$, $PSI = 0$. For example, if the test results are random and the probability of both PPV and NPV is 50%, then the test is useless and $PSI=0$. When $-1 < PSI < 0$ the test is misleading; i.e. the tests results are negatively associated with the true diagnosis. When $\Psi = -1$, the test is always misleading.

The PSI can also be interpreted as the gained probability of correct diagnosis information, i.e. the difference between the joint probabilities of correct

diagnosis (positive or negative diagnosis) $PPV \cdot NPV$ and the joint probabilities of incorrect diagnosis $FPF \cdot FNF$:

$$\begin{aligned} &PPV \cdot NPV - FNF \cdot FPF = \\ &= PPV \cdot NPV - [(1 - PPV) \cdot (1 - NPV)] = \\ &= PPV \cdot NPV - 1 + NPV + \\ &+ PPV - NPV \cdot PPV = \Psi \end{aligned} \quad (\text{Eq. 6})$$

PSI as a difference measure of a diagnostic test

The *PSI* can be interpreted as the difference between the correct prediction of a disease by the test and a false negative result of the test in the target population.

$$PSI = PPV - FNF \quad (\text{Eq. 7})$$

Thus, *PSI* reflects the excess in the proportion of infected people in those with a positive result versus the proportion of infected people when the test is (falsely) negative.

Similarly, one can also interpret *PSI* as

$$PSI = NPV - FPF \quad (\text{Eq. 8})$$

Here, *PSI* reflects the excess in the proportion of uninfected persons when the test yields a negative result versus the proportion of uninfected people when the test is (falsely) positive.

The $NND = 1/PSI$ is analogous to *nns*, to estimate the number of patients who need to be examined in the patient population, in order to correctly diagnose one person (see Table 2). For example, this can be the number of people who would have to undergo a PCR test to correctly diagnose one person. This measure may be abbreviated as the “number needed to diagnose” (*NND*). This information has public health importance. It also enables planning of test services to a specific population,

based on the prevalence of the disease in this population as well as on the technical characteristics of the test.

CALCULATING THE DIAGNOSTIC ACCURACY MEASURES USING THE PREVALENCE AND THE DETECTION ACCURACY MEASURES

The data in Table 1 are usually provided by the manufacturer of a medical test, or obtained in a study with a sample of persons with or without the disease. Such a study does not reflect the prevalence of the disease in the general population because the ratio of the number of persons with the disease to the number of persons without it is artificially determined by the researcher. However, the data in a study that samples persons with positive or negative test results (Table 2) are often unavailable because it is impossible or unethical to follow up persons with negative diagnostic test results. Thus it is necessary to calculate *PPV* and *NPV* from the data in Table 1, using the true prevalence of the disease.

Such calculations can be performed using Bayes’ Theorem (see below for Eq. 9).⁴⁻²¹

Using *sensitivity*=90%, *specificity*=90%, and *Prevalence*=1%, we get *PPV*=8.3%.

A similar equation exists for calculating *NPV*.

Alternatively, a table that reflects the real prevalence of the disease can be constructed, using the sensitivity and specificity, enabling *PPV* and *NPV* calculation (Table 3 in the following example).

Example

The following is an example for calculating *PPV*, *NPV*, *PSI*, and *NND* using data on the prevalence of the disease and given data on the sensitivity and specificity of the test (Table 3).

$$\begin{aligned} PPV &= \frac{\text{sensitivity} \cdot \text{Prevalence}}{\text{sensitivity} \cdot \text{Prevalence} + (1 - \text{specificity})(1 - \text{Prevalence})} \\ &= \frac{\text{sensitivity} \cdot \text{Prevalence}}{\text{sensitivity} \cdot \text{Prevalence} + \text{fpf} \cdot (1 - \text{Prevalence})} \end{aligned} \quad (\text{Eq. 9})$$

Table 3. PPV, NPV, PSI, and NND calculated based on the Prevalence 1%, Sensitivity 90%, and Specificity 90%.

		COVID-19		
		S+ Sick	S- Not Sick	
Test Results	T+ Positive Test	9	99	108
	T- Negative Test	1	891	892
		10	990	1000

$$PPV = 9/108 = 8.3\%, NPV = 891/892 = 99.9\%,$$

$$FPF = 91.7\%, FNF = 0.1\%$$

$$PSI = \Psi = PPV + NPV - 1 = 0.082$$

$$NND = \text{Minimal Number Needed to be tested to diagnose correctly one person} \\ = 1/PSI = 12.2$$

Suppose that a population of 1000 travelers is screened for COVID-19 using the PCR test, and that the prevalence of COVID-19 is 1% (which is an estimated prevalence of asymptomatic persons in Israel who had been exposed to a person with COVID-19). Of the 1000 travelers, only 10 are COVID-19 patients, and 990 are free of the disease. If the sensitivity is 90%, then 9 persons will be correctly diagnosed as infected with COVID-19 (and 1 person will not be diagnosed, false negative). If the specificity is 90%, only 891 persons will be diagnosed correctly as free of COVID-19, but 99 travelers will be falsely diagnosed as having the disease and will be quarantined unnecessarily. The *PPV* is only $9/108=8.3\%$, and the *FPF*=91.7%. Thus, the fraction of false positive diagnoses is very high, and many persons will be quarantined or denied a flight unnecessarily. On the other hand, the *NPV* is 99.9%, and there are only a few false negative diagnoses. The test will be very useful to diagnose person who are not sick with COVID-19, and may be helpful to ensure that passengers are without COVID-19. The

PSI is 8.2%, which is a very low proportion of additional information obtained by the test. However, this is very useful information. Indeed, more than 12 tests will be needed to diagnose one person correctly, because of the high percentage of false positive tests. There were more than a billion United States passengers flying every year before the corona pandemic. Thus, the tests may be very expensive, and will impose significant personal hardship to persons with false positive diagnoses. However, it will be a useful test in order to make sure that no person with COVID-19 will board a flight.

NND with a Changing Prevalence, Where Sensitivity and Specificity Equal 90%

Table 4 summarizes *NND* with a changing prevalence, where sensitivity and specificity are 90%. The *PPV* is increasing and the *NPV* is decreasing with increasing prevalence. When the COVID-19 prevalence is 1%, 5%, or 10% (most estimates in different countries and societies are in this range),

Table 4. Example for *PPV*, *NPV*, *PSI*, and *NND* as a Function of the Prevalence with 90% Sensitivity and Specificity.*

Prevalence (%)	Estimated Prevalence	<i>PPV</i> (%)	<i>FPF</i> (%)	<i>NPV</i> (%)	<i>FNF</i> (%)	<i>PSI</i> (%)	<i>NND=1/PSI</i>
1	Iceland, Israel: Asymptomatic	8.3	91.7	99.9	0.1	8.2	12.2
5	Spain, Israel	33.1	67.9	99.4	0.6	31.6	3.2
10		50.0	50.0	98.8	1.2	48.8	2.1
30	New York City	79.4	20.6	95.5	4.5	74.9	1.3
50	Almost Herd Immunity	90.0	10.0	90.0	10.0	80.0	1.3
70	Herd Immunity	95.5	4.5	79.4	20.6	74.9	1.3
90	Most Infected	98.8	1.2	50.0	50.0	48.8	2.1

* This is an illustrative table. There are no reliable estimates of disease prevalence in various countries.

more than 12, 3, and 2 tests, respectively, are needed to diagnose one person correctly.

DISCUSSION

We showed that using COVID-19 testing to enable re-opening of airports may have its limitations and should be considered with caution. When diagnosing acute infection, it is important to avoid a false positive result that would unjustifiably quarantine a person and all persons that this person has met. It is equally important to avoid false negative diagnoses, because this can falsely reassure actual patients and hinder appropriate contact tracing and isolation. False positives and false negatives may have different implications for public health policy. Indeed, false positive test results may lead an actually healthy person to be wrongfully quarantined or retested, which is of cause inconvenient, infringing on personal freedom of travel, and somewhat costly. However, false negatives are much more damaging to society as a whole, as sick persons, by remaining undetected, can spread the disease in the community or on a plane. These effects should be instrumented into the assessment models, perhaps by adjusting the indices with appropriate (subjective) weights for the *PPV* and *NPV*.

The fraction of errors in diagnoses depends on the technical test characteristics, that is, the sensitivity and specificity, and the prevalence of the diseases in the population of interest. The prevalence can usually be obtained from national serological

tests. Thus, a condition for planning a testing campaign for acute diseases is to use serological tests that estimate the prevalence of past infection.

For example, in the USA there are more than 400,000 people who have had COVID-19. This is approximately 0.1% of the population. Clearly, this is an underestimate of the true prevalence of the disease due to under-testing and under-reporting. If the prevalence of the disease in the USA is 0.1%, then even if the test is very accurate technically and has a specificity and sensitivity of approximately 90%, there is still a very low probability that a positive test is correct (*PPV*=0.9%), and approximately 99.1% of results are false positive. Even a prevalence of 1%, 10 times higher than the official estimate, would yield a *PPV* of 8.3% and 91.7% false positive results, with dire consequences to individuals. However, false negative results, and the risk of onboard infection, would be only 0.1%. Nevertheless, negative tests need to be interpreted with caution, taking into account the pre-test probability of disease (the prevalence) because false negatives may lead to spread of COVID-19.

The high number of tests that need to be performed to diagnose one person correctly, as shown in Table 4, may lead to high costs in any routine screening program.

The attempts to slow down the epidemic and control infections may lead to the use of tests with high sensitivity (and yielding fewer technical false negative results). However, we have shown that in a population with a low prevalence, these tests may

lead to a high percentage of false positive diagnoses and a need for multiple tests at high cost to achieve a correct diagnosis.

We emphasize that in an RT-PCR test the RNA of the virus is detected, but not necessarily an infective virus. Thus, even correct test results may be false positive results, with respect to infectiousness.²⁰ The prevalence of a disease is measured by serological testing. However, it can take up to five days for an infected person to develop antibodies to SARS-CoV-2, and thus the prevalence may be underestimated. Repeated testing can also improve the accuracy of the overall testing program.

We suggest caution in using COVID-19 testing in planning re-opening of airports. It seems prudent to recommend multiple testing of travelers with positive test results, together with obtaining information on previous exposure to persons with COVID-19 or being involved with populations that are known to have high prevalence of COVID-19. Using multiple quick and more accurate tests may be helpful before boarding an airplane.

REFERENCES

1. Johns Hopkins Coronavirus Resource Center. COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Available at: <https://coronavirus.jhu.edu/map.html> (accessed July 28, 2020).
2. Centers for Disease Control and Prevention. Overview of testing for SARS-COV-2. Updated July 2, 2020. Available at: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/testing-overview.html> (accessed July 5, 2020).
3. Sharfstein JM, Becker SJ, Mello MM. Diagnostic testing for the novel coronavirus. *JAMA* 2020;323:1437–8. [CrossRef](#)
4. Watson J, Whiting PF, Brush JE. Interpreting a COVID-19 test result. *BMJ* 2020;369:m1808. [CrossRef](#)
5. Riffenburgh RH. *Statistics in Medicine*. 3rd ed. San Diego, CA: Academic Press; 2012. [CrossRef](#).
6. Feinstein AR. *Principles of Medical Statistics*. Boca Raton, FL: Chapman & Hall/CRC; 2002.
7. Weinstein MC, Fineberg HV, Elstein AS, et al. *Clinical Decision Analysis*. Philadelphia, PA: WB Saunders; 1980.
8. Altman DG. *Practical Statistics for Medical Research*. Boca Raton, FL: Chapman & Hall/CRC; 1991.
9. Sackett DL, Haynes RB, Tugwell P, Guyatt G. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd ed. Boston, MA: Little Brown & Co; 1991.
10. Kraemer HC. *Evaluation of Medical Tests: Objective and Quantitative Guidelines*. London: Sage Publications; 1992.
11. Riegelman RK. *Studying a Study and Testing a Test: How to Read the Medical Evidence*. 4th ed. Philadelphia, PA: Lippincott, Williams & Wilkins; 2000.
12. Knottnerus JA, van Weel C. General Introduction: Evaluation of Diagnostic Procedures. In: Knottnerus JA, ed. *The Evidence Base of Clinical Diagnosis*. London: BMJ Books; 2002:1–18.
13. Sackett D, Haynes RB. The Architecture of Diagnostic Research. In: Knottnerus JA, ed. *The Evidence Base of Clinical Diagnosis*. London: BMJ Books; 2002:19–38.
14. Zhou X-H, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. New York, NY: John Wiley & Sons, Inc.; 2002.
15. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series 28. Oxford: Oxford University Press; 2003.
16. Morabia A, ed. *A History of Epidemiologic Methods and Concepts*. Berlin, Germany: Birkhauser Verlag; 2004.
17. Roman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
18. Grobbee DR, Hoes AW. *Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research*. 2nd ed. Burlington, MA: Jones & Bartlett Learning; 2015.
19. Linn S. A new conceptual approach to teaching the interpretation of clinical tests. *Journal of Statistics Education* 2004;12. Available at: <http://jse.amstat.org/v12n3/linn.html> (accessed July 7, 2020).
20. Linn S. New patient-oriented diagnostic test characteristics analogous to the likelihood ratios conveyed information on trustworthiness. *J Clin Epidemiol* 2005;58:450–7. <https://doi.org/10.1016/j.jclinepi.2004.07.009>
21. Linn S, Grunau DP. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiol Perspect Innov* 2006;3:11. <https://doi.org/10.1186/1742-5573-3-11>