MEDICAL DISPATCH

WHY A NEGATIVE TEST DOESN'T GUARANTEE YOU DON'T HAVE THE CORONAVIRUS

We want coronavirus tests to give us the all-clear. But, in medicine, test results are clues, not answers—and no test is perfect.

By Clayton Dalton

September 16, 2020

Medical tests, including those we use for the coronavirus, are rarely completely definitive and are never infallible. Therefore, a doctor needs to know how often the test is wrong. We have to test the test.
Illustration by Pete Ryan

During my medical residency, I spent a month on the obstetrics ward, learning how to deliver babies. On one occasion, I was paged to a delivery room down the hall. Inside, I saw the mother-to-be propped up in bed, her husband sitting next to her. Then I noticed the crowd. In addition to the obstetricians and delivery nurses, a special team of nurses and pediatricians from the neonatal I.C.U. had gathered.

"Is this a high-risk delivery?" I asked one of the obstetrics residents.

"Maybe," he whispered back. "The baby had a positive screen for cri du chat."

Cri du chat, which means "cat's cry" in French, is a rare genetic syndrome affecting one in fifteen thousand to fifty thousand infants. It was discovered by Jérôme Lejeune, a French geneticist, in 1963. Early in fetal development, genes situated on the short arm of chromosome 5 are accidentally left out, like parcels falling off the back of a delivery truck; affected children can have poor muscle tone, low birth weight, a cleft palate, unusually small heads, speech delays, learning disabilities, and heart problems. As many as one in ten children with cri du chat don't survive the first year of life. The anatomy of the vocal cords is often affected, as well, resulting in a highly abnormal cry that sounds startlingly like the mewing of a cat.

Because the syndrome can cause such significant handicaps, pregnant women with risk factors are often offered a screening test. By sampling fragments of fetal DNA that have migrated from the placenta into the mother's blood, doctors can sometimes detect the fingerprints of a genetic disorder. But, because the test doesn't inspect the infant's DNA directly, it can't provide a perfect picture of fetal genes and chromosomes; it's more like a slightly blurry snapshot. It's possible for a result to be incorrect—falsely positive or falsely negative. If the screening test does come back positive, another test is needed to confirm the diagnosis. This second test is more invasive, and involves taking samples of fetal or placental tissue. And yet, although it's more accurate than the screening test, it's not perfect, either. Even the confirmatory tests can sometimes be misleading.

As we stood near the back of the room, the obstetrics resident told me that the laboring woman had undergone some of these more advanced tests. The results had been reassuring, and the team had told the parents that the fetus was most likely normal. But the tests couldn't say for sure, and neither could we. The neonatal-I.C.U. team was there on standby because, despite our most sophisticated tests, the nature of the child would remain a mystery until the moment of its birth.

In the room, that moment unfolded in stages. First, we glimpsed the child's head; then the shoulders, then limbs, then a body. The room was strung on a wire as the crowd of nurses and doctors watched the delivery in silence. The obstetrician lifted the newborn free of the birth canal. The cord was clamped and cut, and the newborn opened its mouth, filled its lungs, and cried. The exhausted mother beamed as she held her child in her arms. She didn't hear what everyone else heard. The cry was not a normal cry. Shrill and plaintive, it sounded like the mewing of a cat.

Some medical problems are obvious. When the sidewalks in Boston are icy, I see a lot of patients who've lost their footing and fallen onto their outstretched hands; I can usually tell at a glance if they've broken a wrist. But most medical problems aren't obvious. They arise from hidden processes that occur within the body, in tissues, cells, enzymes, or genes. They manifest only indirectly, through symptoms or signs. As the American Medical Association noted, in 1912, internal medicine is concerned primarily with "abstract problems" and the "intangible struggle against unseen infections." Choosing an appropriate treatment depends on discerning the cause of an illness—and yet there are a number of possible causes for most symptoms. How do doctors connect a symptom to a cause?

The answer, of course, is that we test. To test is to examine something critically, to put it to the proof. The word is often thought to derive from the Latin *testari*, meaning to testify. But the Reverend Abram Smythe Palmer, a respected nineteenth-century lexicographer, placed its origins with the Old French *test*, which referred to a vessel used for cupellation, the extraction of precious metals with heat. " '*To test*' a thing," Palmer wrote, in 1882, "is properly to submit it to the crucible or melting pot, to assay the quality of its metal." The emergency department often feels like a crucible, where we approach our patients, undifferentiated, as a minter approaches ore. We apply our tests to find out what lies within.

Today, amid the coronavirus pandemic, we are thinking about medical tests more than usual. Often, we have a fairly simple vision of how tests work. We picture them as high-tech and definitive; we see them cutting through the ambiguities of an often asymptomatic virus. We hope that, by speedily distinguishing between the sick and the well, tests might help us establish defensive cordons around schools, workplaces, and public events. Colleges and universities have used coronavirus tests to sort students into different dormitories as they return to campus. Sports teams, too, have created testing-based "bubbles" within which they hope something like normal life can go on. We envision simple steps—a nasal swab, a sample tube, an expensive machine—followed by bad news or an all-clear.

But physicians tend to approach testing more cautiously, and in an incremental fashion. In fact, we are always testing, often in ways that don't involve technology. One of our most important tests is one of our simplest: the visual assessment, what we call "eyeballing" a patient. There's a double meaning to the statement "the doctor will see you now"; just laying eyes on someone can yield a huge amount of information. We can quickly tell whether a patient is critically ill or stable; we can often recognize critical illness from the way someone looks in a doorway. We may not know the cause, but we can sense the severity.

A lot of testing happens through language. Around 100 A.D., Rufus of Ephesus, a Greek physician, published the first treatise on taking a medical history; he described aspects of the patient interview that medical students still learn today, such as asking about the location, duration, and character of pain. I learned many of these principles in medical school but didn't realize until my residency that interviewing patients is actually a way of testing them. "Think about it this way," one of my supervisors said. "When you question a patient about their symptoms, do their answers influence your suspicion of potential causes?" They do, just as the results of a blood test would.

In the centuries after Rufus, doctors pioneered new ways of testing the body. "Water casting," or inspecting urine, became the diagnostic test of choice in medieval Europe. The Jerusalem Code of 1090 made doctors liable to public beatings if they failed to examine it. A blood-pressure measurement was taken for the first time in 1733, when an English clergyman inserted a brass pipe into the artery of a horse (he found that the animal's blood pressure rose by a factor of four when it began struggling). In the seventeen-fifties, Leopold Auenbrugger, an Austrian physician, developed a groundbreaking technique called percussion, which is still in use today. After observing his father tapping wine casks to determine how full they were, he realized that a similar method could be employed to localize diseases, such as pneumonia, within the body of a living patient. He discovered that a healthy lung, when rapped with a couple fingers, sounded like "a drum covered with a thick woolen cloth," whereas a diseased region was "entirely destitute of the natural sounds."

Within a hundred and fifty years, the stethoscope, the thermometer, the blood-pressure cuff, and the electrocardiogram became standard medical tools. Techniques to measure blood-clotting time and the concentration of white blood cells became available; diagnostic tests for tuberculosis, diphtheria, cholera, and typhoid were introduced. But the most astonishing addition to the arsenal of medical tests during this period was the X-ray, developed in 1895. A medical reference book that I keep on my desk, published in 1910, is ecstatic in its description. "Verily the X-ray opens the field for the grandest of electrical possibilities," the authors write, allowing physicians to make "far-reaching diagnoses, and to ascertain with certainty the whole internal structure of the human body." It must have seemed that this flood of new technology would bring doctors closer to certainty than ever before. But it was the X-ray which first demonstrated that testing was much more complicated than anyone realized and that medical tests might deceive as well as diagnose.

In an essay for the *Harvard Law Review*, in 1897, Oliver Wendell Holmes, Jr., wrote about an eminent judge who refused to issue verdicts until he was absolutely certain that they were correct. It was an understandable impulse, Holmes argued, given the "longing for certainty and for repose which is in every human mind." But he thought it misguided. "Certainty generally is illusion," Holmes concluded, "and repose is not the destiny of man."

Holmes was right that a longing for certainty is fundamental to human psychology. Studies have suggested that we tolerate physical pain better than uncertainty; the psychologist Jerome Kagan has argued that the impulse to resolve uncertainty is a chief driver of human behavior. Uncertainty about the world around us is difficult enough. But when it comes to our own bodies it can be especially unsettling, and the impulse to peer inside and resolve it is powerful. From 2001 to 2002, researchers asked five hundred Americans whether they would prefer a thousand dollars in cash or a free full-body CT scan; seventy-three per cent chose the scan, even though they had no symptoms of disease at the time. In the E.R., it's common for my patients to press for an X-ray or scan, even if my assessment suggests that they don't need one. I imagine it's because they find the idea of a test to be more definitive, more objective, than my words of reassurance.

Research supports the idea that many patients view medical tests as more definitive than they really are. In 2006, a survey conducted in Germany found that nearly half of all respondents believed that the results of a mammogram would be "absolutely certain," when in fact mammograms may be wrong as often as one in five times. Recently, I overheard a woman telling her companion about a nephew who seemed to be exhibiting symptoms of bipolar disorder; he refused to "get tested for it," she complained, as though a simple test could establish the source of his trouble. I've noticed a similar attitude about diagnostic tests for the coronavirus. Last month, a few family members came down with fevers and coughs but tested negative for the virus. "I guess they didn't have it, after all," someone said, relieved. The reality is that medical tests, including the ones we use for the virus, are rarely completely definitive and are never infallible; every test is susceptible to error. In order to interpret a test result, therefore, a doctor needs to know how often the test is wrong. We have to test the test.

Testing the test was Jacob Yerushalmy's idea. Born near Jerusalem, in 1904, Yerushalmy emigrated to the United States to study mathematics at Johns Hopkins, where he became interested in the nascent field of biostatistics; he made important contributions in public health before arriving at the University of California, Berkeley, in the late nineteen-forties, where he helped establish the country's first Ph.D. program in the subject. In 1947, he published what would become his most influential paper. Yerushalmy set out to compare different X-ray techniques for diagnosing tuberculosis. Each patient in his study had four different types of chest X-rays taken. Five expert radiologists then read the films, identifying the patients they thought had tuberculosis in their lungs. Yerushalmy already knew that every test was vulnerable to error. "Even with the best techniques," he wrote, "it is not always possible to raise the level of diagnosis to absolute certainty." But the results of his study were surprising. No single X-ray technique emerged as clearly superior; different techniques performed better for different readers, and performance varied by as much as thirty per cent between techniques. To make matters worse, the readers themselves turned out to be unreliable. Unbeknownst to them, Yerushalmy had asked each reader to interpret each film twice. There was an unsettling degree of inconsistency—one reader contradicted himself more than twenty per cent of the time.

Yerushalmy concluded that these limitations were inherent in the technology itself and in the subjective nature of interpretation. There was no getting around them. His breakthrough was in realizing that he could quantify the uncertainty: if he could put a number on it, doctors could have a much more accurate sense of how much they could trust a test. Yerushalmy developed two important metrics that—despite their brain-teasingly similar names—have since become the standard by which all tests are evaluated. The first, sensitivity, describes a test's ability to correctly identify patients who have the disease in question. (If a test detects ninety-five out of a hundred patients who have the condition, its sensitivity will be ninety-five per cent.) The second, specificity, describes a test's ability to avoid confusing other conditions for the one it's supposed to detect. (If a test clears ninety-three out of a hundred people who are disease-free, its specificity is ninety-three per cent.) A test with low sensitivity is more likely to deliver a false negative. A test with low specificity is more likely to deliver a false positive. Together, sensitivity and specificity suggest how much, and in what ways, doctors should trust a test.

Sensitivity and specificity vary from test to test, and are influenced by a variety of factors. Some relate to human error: a person can misinterpret an image or process a sample incorrectly. Others have to do with tests themselves: a chemical analysis might produce unavoidable, random errors, or an MRI or CT scanner might generate small blemishes on its images. Diagnostic tests for the coronavirus mostly rely on a technology called polymerase chain reaction (PCR), which detects the virus's genetic material. Because of how such tests work, they are highly specific but only moderately sensitive. A positive PCR test for the coronavirus means that the virus is very likely present; on the other hand, there's a decent chance that a negative test could be wrong. Although there is disagreement about how accurate

PCR tests for the coronavirus are, current estimates suggest that, in the field, the false-negative rate could range as high as thirty per cent. It's possible for improper swab collection or storage to decrease the quantity of virus in each sample; especially early in the course of an infection, an infected patient may have viral levels below the "detection threshold" of the test. Some of these factors can be mitigated, but not all.

As Yerushalmy considered how to quantify test error, he uncovered an even thornier problem. In order to test a test, you need to have another way of identifying positive patients—a standard by which you can evaluate the test under review. (How else will you know if your test has missed someone?) Creating such a baseline, of course, usually requires another test. "It is, in a sense, a vicious circle," Yerushalmy wrote. Someone has to test the baseline test; testing that requires another test, and so on. In his study of X-rays, Yerushalmy tried to sidestep this problem by using a consensus of several radiologists as a baseline. He had no way of testing the accuracy of their consensus, but it was the best he could do. Today, a similar approach is being used to evaluate tests for the coronavirus. According to a recent article in the *BMJ*—a medical journal published in Britain since 1840—evaluators are judging the accuracy of tests by comparing their results with everything that's known about a patient, including symptoms, basic blood tests, imaging studies, and repeated coronavirus tests. An imperfect test is being used to help verify itself. This recursivity, which is not uncommon in medical testing, is part of why it's so hard to make tests perfectly reliable.

The fact that all tests are fallible to varying degrees means that, for many diseases, we must grow comfortable with treating diagnosis as a probability rather than a yes-or-no answer. Instead of asking whether we have COVID-19, or any other illness, we should ask what the likelihood is that we have it. There will always be some degree of uncertainty about what's happening inside our bodies. In the absence of certainty, probability is the next best thing.

Probability theory has its roots in Islamic studies of cryptography made in the eighth century; in the seventeenth, two mathematicians gave the theory its modern form. Pierre de Fermat and Blaise Pascal were both fascinated by games of chance, and together they worked out a mathematical approach for describing the likelihood of different rolls of dice. It was one of the first attempts to measure uncertainty in concrete terms, and it turns out to be extraordinarily useful in medical diagnosis, too. For doctors, one of the most important probabilistic concepts is the "base rate." In his book "Thinking, Fast and Slow," the psychologist Daniel Kahneman uses a word problem to explain it:

Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail. Is Steve more likely to be a librarian or a farmer?

Most readers choose librarian because Steve's description so closely fits our idea of what librarians are like. But, in fact, Steve is much more likely to be a farmer, because male farmers in the United States outnumber male librarians by a factor of twenty. If there are twenty times more male farmers than male librarians, your chances of finding a male farmer—even a very shy one—are much greater than your chances of finding a male librarian. This relative prevalence is known as the base rate, and it exerts a profound influence on the probability of an outcome.

In medicine, the base rate reflects how common a disease is in a population. It can have surprising and sometimes extreme effects on the reliability of a test result. Imagine administering a test with eighty-per-cent sensitivity and specificity to a population in which one person in a thousand has the disease you're looking for. If a perfectly accurate version of this test were administered to a thousand people, it would deliver one positive result. But, because the test's specificity is eighty per cent, it will deliver a false positive twenty per cent of the time—resulting in two hundred false positives for every true positive. The problem is that the tendency to produce false positives is more or less fixed; it's inherent in the test. The number of true positives, however, depends on how common the disease is. The more uncommon it becomes, the more false positives outnumber true ones.

The base-rate effect is why more testing isn't always better and why medical authorities sometimes recommend against the routine use of certain tests, such as the PSA test for prostate cancer or mammograms for breast cancer. Those are among the most common cancers; even so, the proportion of the population that has them is small. Their low base rates mean that a positive test is more likely to be a false alarm. The chance that a positive PSA test for prostate cancer represents a true positive, for example, is only about thirty per cent; the false-positive rate for mammography varies from less than five per cent to as much as fifty per cent, depending on the skill of the radiologist.

Antibody tests for the coronavirus suffer from the same base-rate problem; it's why the Centers for Disease Control and Prevention and the Food and Drug Administration have issued so many cautionary statements about their use. Current estimates suggest that, in the different regions of the United States, between five and twenty-five per cent of the population has been exposed to the virus. If you're testing a population with a base rate as low as five per cent, even a test with ninety-per-cent sensitivity and specificity will still miss half of all cases, and more than half of all positive results will be false positives. (By contrast, if the base rate rises to fifty per cent, the false-positive rate plummets to less than five per cent.) Here, we encounter another vicious circle. Many researchers are using antibody tests to find out how many people have been exposed to the virus. But, to judge the accuracy of those tests, researchers must make a calculation in which the base rate itself is a variable. In this way, tracking a virus as it spreads across a population can feel like stepping into an Escher print.

The base rate is an important factor when we judge the likelihood of any one person having a disease. But many other factors shift the probability, too. When I evaluate a patient in the E.R., every single bit of data I gather, consciously or not, is assimilated into a mental calculus that moves me toward a likely diagnosis. If you come to the hospital with chest pain, I consider your age, gender, and ethnicity; I learn about your other medical problems; I inquire into your habits—do you smoke?—and your parents' heart histories; I listen closely as you describe the pain, and ask a few questions about your symptoms. Some informal tests at the bedside, including a visual assessment and physical exam, refine the calculus further. Combined with an understanding of the base rates of various conditions, this nuanced shifting of weights and measures gives me a sense, before any further testing, of the probability that you could be having a heart attack or any other illness. A test result doesn't replace that probability; it shifts it.

The framework for this kind of reasoning, in which new information is used to alter the probability of a diagnosis, was created in the mid-eighteenth century by Thomas Bayes, an English Presbyterian minister and mathematician. Bayes developed a mathematical formula that, when applied to medical decision-making, takes into account the sensitivity and specificity of a test, using them to modify the pre-test probability of a disease. Suppose that a person coming into the E.R. with chest pain is a man. He says that the pain radiates all the way to his jaw. He's a smoker and has several other risk factors for cardiac disease. The first step in using Bayes's formula is to assign a number to my pre-test suspicions. Everything I know tells me that there's a high probability he's had a heart attack. Perhaps a reasonable likelihood is eighty-five per cent.

Now I order some tests, starting with an EKG. Suppose it comes back normal. Does that mean that the man didn't have a heart attack? Bayes's formula allows me to plug in my pre-test probability—eighty-five per cent—along with the sensitivity and specificity of the EKG, which can miss as many as thirty per cent of major heart attacks. The formula tells me that, even with a normal EKG, the probability that my patient has had a heart attack is still greater than sixty per cent —high enough that it makes sense to treat him as a heart-attack victim, even though I can't know for sure what happened.

Such probabilistic diagnoses can be hard to accept. Recently, I took care of an elderly woman who was experiencing confusion and general weakness. When we ordered an MRI of her brain, the images displayed a small abnormality, which could have been either a small stroke or a random blip generated by the machine. My pre-test sense, based on her symptoms, was that the probability of stroke was low; the images weren't compelling enough to shift that probability in a significant way. When I shared this diagnosis with her family, they weren't comfortable with the idea that, although she probably hadn't had a stroke, we couldn't say for sure. They wanted to know definitively, one way or the other.

And yet, for those making decisions based on diagnoses, it's important to keep uncertainty in view. Colleges and universities across America have begun testing all students upon arrival; the plan is often to isolate students who have tested positive from those who have tested negative. But we know that, for a variety of reasons, the PCR tests used to detect the coronavirus can miss a sizable number of cases. The virus is so contagious that even a small number of students who have received false negatives can spread it widely. This may be one of the reasons that even colleges with comprehensive testing strategies, such as the University of Illinois at Urbana-Champaign, are now struggling to contain serious outbreaks on campus.

The fact that tests aren't perfect has tempted some colleges to deëmphasize them. Officials at the University of North Carolina at Chapel Hill have defended their decision to test only students who show symptoms or have been exposed to those known to have the virus by claiming that universal testing would create a "false sense of security." They're right that testing isn't a panacea, but their strategy will invariably miss asymptomatic or pre-symptomatic cases. Northeastern University, in Boston, might have a better approach. Recently, it made headlines after it dismissed eleven students for violating the school's social-distancing policies. Although the move might seem Draconian, Northeastern's emphasis on enforcing additional measures to limit transmission of the virus is sensible. The university tests all faculty and staff twice a week, and all incoming students on their first, third, and fifth days on campus; students can only attend classes in person if all three tests are negative. But Northeastern's administration knows that a negative coronavirus test can't be treated as a get-out-of-jail-free card. Just as many factors contribute to the likelihood of a positive diagnosis, so many simultaneous interventions—not just a testing system—need to be put into place to limit spread.

The complexities of medical testing shouldn't just influence how institutions use tests. They have implications for all of us, as individuals, as we navigate the pandemic. Suppose that a man develops a fever and cough. He lives in a city with rising coronavirus cases. His roommate recently tested positive for the virus. As he tells me all this, my suspicion that he has covid-19 climbs higher: I estimate it at seventy-five per cent. Then he gets a test, and it comes back negative. Plug the most cautious estimates about the false-negative rates for PCR tests into Bayes's formula, and you'll find that, even with a negative result, his post-test probability for having the virus is fifty per cent.

This man was my youngest brother. He wanted to know whether he should isolate himself from our father, a two-time cancer survivor. I told him that, even though he'd tested negative, he should act as though he had the virus. He was surprised to learn how much uncertainty remained despite his negative result. He agreed to keep his distance from Dad.

MORE MEDICAL DISPATCHES

- Surviving a severe coronavirus infection is hard. So is recovering.
- Some hospitals have postponed cancer surgeries because of the coronavirus crisis. How do doctors assess urgency during a pandemic?
- It is not too late to go on the offense against the coronavirus. This five-part public-health plan may be the key.
- The loneliness and solidarity of treating coronavirus patients in New York.
- To fill the vacuum left by the federal government, doctors are relying on informal networks to get the information and support they need.
- Conflict and confusion reign at New York hospitals over how to handle childbirth during the pandemic.
- In countries where the rate of infection threatens to outstrip the capacity of the health system, doctors are confronting ethical quandaries for which nothing in their training prepared them.

Clayton Dalton is a resident physician at Massachusetts General and Brigham & Women's Hospitals, in Boston.

More: Coronavirus Medicine Testing Health Public Health

THE DAILY

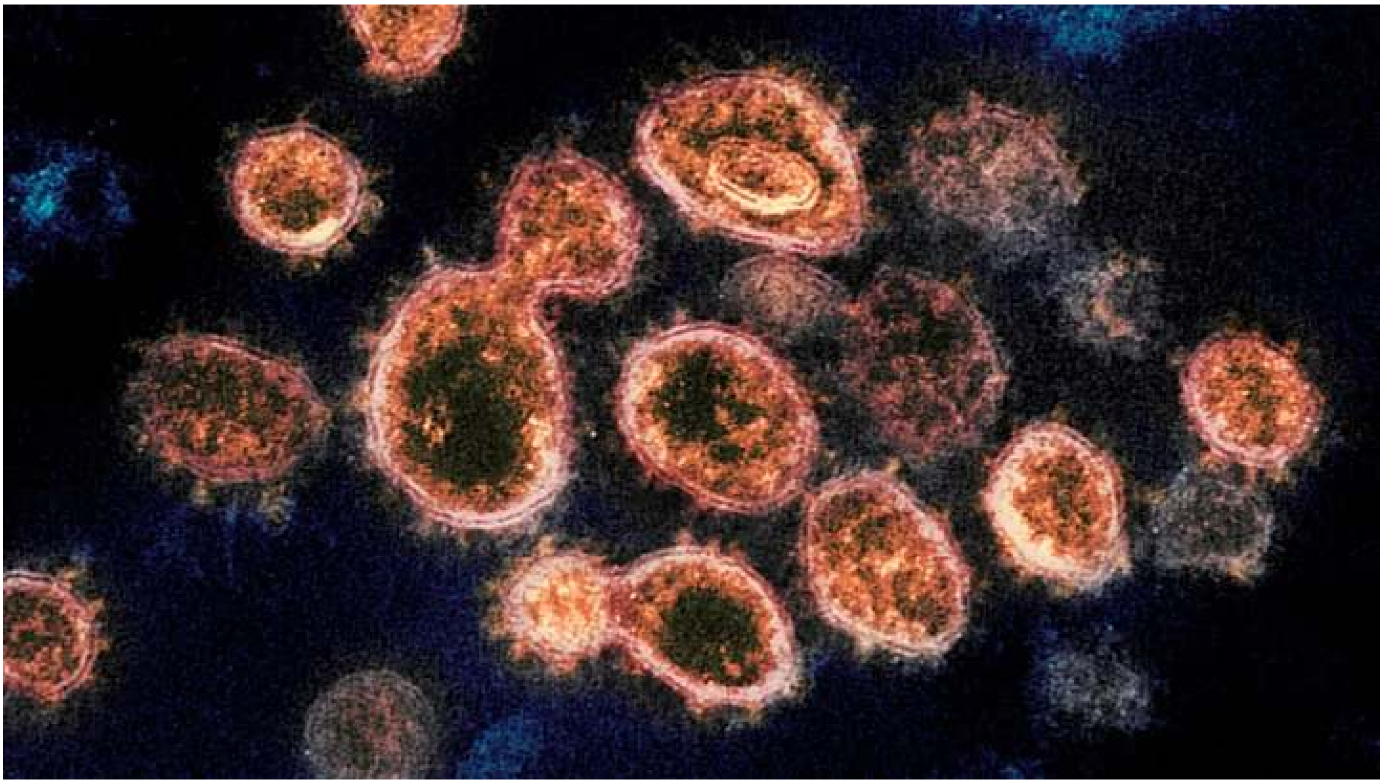
Read More

MEDICAL DISPATCH

THE RISKS OF NORMALIZING THE CORONAVIRUS

What do we lose when we become numb to mass death?

By Clayton Dalton



MEDICAL DISPATCH

THE LONG GAME OF CORONAVIRUS RESEARCH

Warp-speed vaccine trials grab our attention, but more deliberate work is just as urgent.

By Jerome Groopman

VIDEO

A DAUGHTER FORCED TO SAY GOODBYE OVER A VIDEO CALL

Reflecting on her family's firsthand experience with the coronavirus, a New York nurse illuminates the personal tragedy of the disease's toll.

MEDICAL DISPATCH

THE COMPLICATED ETHICS OF KEEPING A COVID-19 PATIENT BREATHING

Normally, a tracheostomy is a straightforward and uncontroversial procedure. During the pandemic, the question of whom to trach, and when, has presented doctors with one of their most difficult decisions.

By Zach Helfand

THE NEW YORKER

SECTIONS

News

Books & Culture

Fiction & Poetry

Humor & Cartoons

Magazine

Crossword

Video

Podcasts

Archive

Goings On

MORE

Customer Care

Buy Covers and Cartoons

Digital Access

Newsletters

Jigsaw Puzzle

RSS

Site Map

About

Careers

Contact

FAQ

Media Kit

Press

Accessibility Help

Condé Nast Spotlight