PROFESSOR: One of the most important applications of conditional probability is in analyzing the results of diagnostic tests of uncertain reliability. So let's look at a fundamental example.

Suppose that I have a diagnostic test for tuberculosis. It really sounds great because it's going to be 99% accurate-- in fact, more than 99% accurate, really, because here are the properties this test has. If you have TB, this test is guaranteed to detect it and say, yes, you have TB. If you don't have TB, 99% of the time, the test says correctly that you don't have TB, and 1% of the time, it gets it wrong.

Now, suppose the doctor gives you the test and the test comes up saying that you have TB. That's kind of scary because TB is, in fact, quite a serious disease. It's getting worse because there are all of these antibiotic-resistant versions of TB. Now in Asia, where all the known antibiotics are not very effective-- if effective at all-- of curing it, and this test that was 99% accurate says I have this disease, it sounds really worrisome.

But in fact, we can ask more technically, should you really be worried? What is the probability given that this apparently highly accurate test says you have TB? What's the probability that you actually have TB? That's what we want to calculate. What's the probability that you have it?

So in other words, I want the conditional probability that I have TB given that the test comes in positive. The test says, yes, you have TB. That test positive is a big word that I won't have room for on other slides, so let's just abbreviate it by [? plus. ?] Plus means, in green, that the test said, yes, positive-- you have TB. OK, so that's the probability that we're trying to calculate, this conditional probability.

What do we know about the test? Let's translate the information we have about the test into the language of conditional probability. And the first thing we said was that the test is guaranteed to get it right if you have TB. So given that you have TB, the probability that the test will say so-- it will return a positive result-- is 1.

Given that you don't have TB, the probability that the test will say that you do have TB is only 1 in 100. Because 99% of the time, it correctly says you don't have TB. And 1% of the time, it says oops, you do have TB.

So this is what's called a false positive rate. It's falsely claiming that you have TB when you really don't. And that rate, we're hypothesizing, is only 1%.

Now, what we're trying to calculate, again, is the probability that you have TB given that the test came in positive and said you had TB. Well, let's look at the definition of conditional probability. The probability that you have TB given that the test came in positive, that said you do, is simply the probability that both the test comes in positive and you have TB divided by the probability that the test comes in positive.

Well, using the definition of conditional probability again, this intersection, this AND of having TB and the test coming in positive, is simply the probability that the test comes in positive given that you have TB times the probability that you have TB. Now, this one we know. It's 1 because the test is perfect. If you have TB, the test is definitely going to say positive. So that lets me simplify things nicely.

What I've just figured out is the probability that you have TB given that the test says you do is simply the quotient of the probability that you have TB given no other information and the probability that the test comes in positive. Well, what is that probability that the test comes in positive? How are we going to calculate that? That's the key unknown here.

And we're going to use the probability rule, the total probability rule. Total probability says that you do or you don't have TB. So that the way to calculate the probability that the test comes in positive is to look at the probability that the test comes in positive when you do and don't have TB. And we know those numbers.

So let's look at the total probability formula. The probability that the test comes in positive is simply the probability that it comes in positive if you have TB times the probability you have TB, plus the probability it comes in positive given that you don't have TB times the probability you don't have TB.

Well, we know a lot of these terms. Let's work them out. Well, the probability the test comes in positive given that you have TB is 1. And the probability that the test comes in positive given that you don't have TB is 1/100. That's the false positive rate. We figured that already.

What about the probability that you don't have TB? Well, that's simply 1 minus the probability that you do have TB. Now I have this nice arithmetic formula in the probability of TB. So I wind up with the probability of TB plus 1/100 minus [? 1/100 ?] of the probability of TB.

It leaves me with 1/100 plus the 99/100 of TB. So that's what this simplifies to. The probability that the test comes in positive given no other information is 99/100 of the probability that a person has TB plus 1/100. We'll come back to this formula.

Well, we were working on the probability that you have TB given the test came in positive. We figured out that it was this quotient. And now, I know what the denominator is. The denominator is 99/100 times the probability of TB plus 1/100. Multiply numerator and denominator through by 100, and you get that the probability that you have TB given that the test says you do is 100 times the probability that you have TB divided by 99 times the probability that you have TB plus 1.

So let's hold formula. Notice the key unknown here is the probability that you have TB independent of the test, the probability that a random person in the population has TB. If we can figure that out or if we can look that up, then we know what this unknown is, the probability have TB given that the test says you do.

Well, what is the probability that a random person has TB? Well, there were 11,000 cases of TB reported in 2011, according to the Center for Disease Control in the United States. And you can assume that there's going to be a lot of unreported cases if there are 11,000 reported ones, because a lot of people don't even know they have the disease.

So let's estimate, on that basis given that the population of the US is around 350 million, that the probability of TB is about 1/10,000. Let's plug that into our formula. The probability that you have TB given the test is positive is this formula. When I plug in 1/10,000 for TB. I get 100/10,000 and 99/10,000 plus 1.

Well now, I can see that the denominator is essentially 1. It's 1.01. And the numerator is 1/100. And this is basically about 1/100.

In other words, it's not very likely that you have TB. Because of the relatively high false positive rate that was relatively high of 1%, that false positive rate washed out the actual number of TB cases, which the TB rate was only 0.01%, so that almost all of the reports of TB were caused by the high false positive rate. And that means that when you have a report that you've got TB, you still only have a 1% chance that you actually have the TB.

So the 99% accurate test was not very useful here for you to figure out what kind of action to take and what kind of medicine to take or treatment to take given that the test came in

positive. With 1 in 100 chance, the odds are you won't do anything, in which case you can wonder why your doctor gave you the test.

Well, the 99% test sounds good. We figured out that it isn't. And a hint about why 99% accurate isn't really so useful is that there's an obvious test that's 99.99% accurate. What's the test? Always say no. After all, the probability is only 1 in 10,000 that you're going to be wrong. And that's the 99.99% rate.

So it sounds as though this test is really worthless. But no, it's not. If you think about it a little bit, it will be useful. And I'll explain that in a minute. I forgot I'm getting ahead of myself. Because the basic formula that we used here was we figured out what the probability of TB given that the test said you had TB in terms of the inverse probabilities which we knew-- that is, the probability that the test came in positive given that you had TB.

This is an example of what's a famous rule in probability theory. It's called Bayes' rule, or Bayes' law. And this is it. It's just stated in terms of arbitrary events A and B. It expresses the probability of B given A in terms of the probability of A given B and the probabilities of A and B independently.

Now, I can actually never remember this law, but I re-derive it right every time I need to do it as we've done in the previous slides. It's really a quite straightforward law to derive and prove.

But let's go back to this 99% accurate test that seemed worthless since there was a trivial test that was 99.9% accurate. But in fact, it's really helpful because it did increase the probability that you had TB by a factor of 100. Before you took the test and before you know anything, you thought that your probability was the same as everybody elses-- about 1 in 10,000. Now the test says the probability that you have TB is 1 in 100. That's a hundred times larger.

What's the value of that? Well, suppose you only had 5 million doses of medicine to treat this American population of 350 million people. Who should you medicate? Well, if you medicated a random 5 million people out of 350 million, the likelihood that you're going to get very many of the real TB cases is small. It's only going to be about 1 in 30. You'll only get about 1/30 of the cases.

But if you use your 5 million doses to medicate the 3.5 million people who would test positive under this 99% accurate test, then when you test all 350 million people, you're going to get about 3.5 million who test positive. You have enough medication to treat all of them. And if you

treat all of them, you're almost certain to get all of the actual TB cases, all 10,000 of them.

So the 99% accurate test does have an important use in this final setting, a lot more so then the 99.99% accurate test that simply always said no-- food for thought.