Grade 7/8 Math Circles
March 4 & 5 & 6 & 7, 2024
The Bayesian Clinical Diagnostic Model

Medical Tests

No medical test is 100% accurate. Let's say that you went to the doctor's office to get a test that reported on whether or not you have a disease. You could get a **positive test result** (i.e., the test indicates that you have the disease) but not actually have the disease. Or you could get a **negative test result** (i.e., the test indicates that you don't have the disease) and actually have the disease.

An example of a medical diagnostic test that you could do at home is the SARS-CoV-2 rapid antigen test, which tests for the presence of COVID-19. If you had a positive test result, what are the chances that you had COVID-19? If you had a negative test result, what are the chances that you had COVID-19? Would these probabilities change if you were asymptomatic or under the age of 35? These are some of the important questions that require the use of mathematics, statistics, and probability to answer.

Here are some examples of medical tests that could indicate the presence or absence of a disease:

- Cancer screening
- Prenatal ultrasounds to detect birth defects
- Testing for infections (strep throat, measles, etc.)

Sensitivity and Specificity

Imagine that you actually had a certain disease. Then, the chance that a test correctly produces a positive test result is called the **sensitivity** of the test. For example, a test with a sensitivity of 75% would correctly report a positive test result three quarters of the time in people who definitely have the disease.

Now imagine that you didn't have the disease. Then, the chance that a test correctly produces a negative test result is called the **specificity** of the test. For example, a test with a specificity of 25% would correctly report a negative test result one out of four times in people who definitely don't have the disease. (The rest of the time, the test would incorrectly report a positive test result.)

Here are the terms we have introduced so far:

- **Positive test result:** The test indicates that you may have a particular disease.
- **Negative test result:** The test indicates that you might be free from a particular disease.
- **Sensitivity:** The probability that the test produces a positive test result, given that one actually has the disease. This is also known as the true positive rate (TPR).
- **Specificity:** The probability that the test produces a negative test result, given that one actually does not have the disease. This is also known as the false positive rate (FPR).

We can shorten "positive test result" to just "positive." Then we can distinguish between a *false positive* and a *true positive*. Can you guess what these terms mean?

- A false positive is a positive test result for someone who does not actually have the disease or condition.
- A true positive is a positive test result for someone who actually has the disease or condition.

Similarly,

- A false negative is a negative test result for someone who actually has the the condition.
- A true negative is a negative test result for someone who does not actually have the disease or condition.

So you can think of a false result as a <u>wrong test result</u> and a true result as a <u>correct test result</u>.

The more true results that a test produces (true positives and true negatives), the better. The **accuracy** of the test is the proportion of true results to all results that a test produces.

Example 1

(a) If you have a false positive test result, do you actually have the disease?

(b) If you have a true negative test result, do you actually have the disease?

(c) If you have a false negative test result, do you actually have the disease?

(d) If you have a true positive test result, do you actually have the disease?

Solution: (a) No. (b) No. (c) Yes. (d) Yes.

Example 2

When a group of 20 patients with confirmed COVID-19 is tested using a rapid antigen test, 15 of the patients have a positive test result and the remaining 5 have a negative test result. In a group of 3 individuals who are confirmed to not have COVID-19, all three of them have a negative test result. What is the sensitivity and the specificity of the test based on these results?

Solution: The sensitivity of the test is the probability that someone with COVID-19 tests positive. The specificity of the test is the probability that someone without COVID-19 tests negative. Based on the information provided in the problem, we have

$$\begin{aligned} \text{sensitivity} &= \frac{TP}{\text{everyone who actually has the disease}} = \frac{15}{20} = 75\% \\ \text{specificity} &= \frac{TN}{\text{everyone who doesn't have the disease}} = \frac{3}{3} = 100\%. \end{aligned}$$

Example 3

(a) Suppose that a test for a certain condition has a sensitivity of 100% and a specificity of 95%. Give an exact value for the number of false negatives.

(b) Suppose that a test for a certain condition has a sensitivity of 95% and a specificity of 100%. Give an exact value for the number of false positives.

(c) Suppose that a test for a certain condition has a sensitivity of 82% and a specificity of 100%. Out of 1000 people tested, 50 test positive and the rest test negative. Pick the correct statement below.

 i. The number of true positives is exactly 50.

 ii. The number of true positives is less than 50.

(d) Suppose that a test for a certain condition has a sensitivity of 100% and a specificity of 95%. Out of 1000 people tested, 50 test positive and the rest test negative. Pick the correct statement below.

 i. The number of true positives is exactly 50.

 ii. The number of true positives is less than 50.

Solution:

(a) Someone who actually has the disease can receive either a true positive or a false negative test result. Recall that the sensitivity is the probability that someone with the disease receives a true positive result. If the sensitivity is 100%, then the number of false negatives is 0.

(b) Someone who doesn't have the disease can receive either a true negative or a false positive test result. Recall that the specificity is the probability that someone with the disease receives a true negative result. If the specificity is 100%, then the number of false positives is 0.

(c) The specificity is 100%, so just like in part (b), we know that there are no false positives. Therefore, option i. is correct: the number of true positives is exactly 50.

(d) The specificity is less than 100%, so there are some false positives out of the 1000 people tested. Therefore, option ii. is correct: the number of true positives is less than 50. **Bonus:** Since the sensitivity is 0%, the exact number of false negatives is 0. Since the specificity is 95%, then 950 represents 95% of the total number of true negatives and false positives, so there are 50 false positives. Since this adds up to 1000, then there are zero true positives.

This section is optional reading and not covered during the Math Circles session.

Statistical Populations

In statistics, a *population* is a group of people, things, or events that is of interest to some question or analysis. When dealing with diagnostic tests, the population is always a group of people who either have a certain disease/condition or not.

Target and Study Populations

It is important to be precise and descriptive when defining a population. Consider the following two populations and their descriptions.

Population 1: All teenagers (ages 13-19).
Population 2: All teenagers (ages 13-19) who have had, currently have, or will have a COVID-19 infection.

The first population is probably harder for us to make reliable conclusions about. The cost of carrying out a study on all teenagers on Earth is way too high. (We may have to travel to secluded parts of the world whose culture and health conditions differ vastly from ours, for example.) Moreover, is this even the population we want to make conclusions about? Public health authorities are often only responsible for a smaller and more localized jurisdiction—such as a city, province, or country. They may also be interested with the effects of a disease or preventative measure, rather than everyone. For example, they may be interested in studying the effects of a certain vaccination or the effects of a certain disease.

The second population is much more precise and could be more helpful. This can be an example of a population we want to consider when studying the prevalence of long COVID. When studying how frequently long COVID occurs in teenagers, we are only considering people who has had COVID-19.

Even though we may want the results of our study to apply to all teens who have had, currently have, or will have a COVID-19 infection (Population 2), we often cannot conduct a study that includes everyone in this population. The **target population** is the population that we want our results to apply to (i.e., Population 2). The **study population**, on the other hand, is the smaller population that can actually be in our study.

Here's an example of a study population for our study on the prevalence of long COVID.

Population 3: Participants aged 13-19 enrolled in the XYZ study between March 31, 2020 and August 2, 2021, who did not meet the exclusion criteria.

Perhaps the XYZ study is based in the Netherlands and is 98% white and 58% female. If these people are of a higher socioeconomic status status than most teenagers—and this significantly reduces the chance of long COVID, then there would be **study error**. Determining the presence of study error requires the consultation of subject-matter experts, and goes beyond the realm of the mathematics. However, it is critically important in any statistical study to precisely define and clearly communicate what the target and study populations are.

To recap:

- **Target population** is the population that we want our results to apply to.
- **Study population** is the population that actually has the potential to be included in our study (i.e., who we *could* study)
- **Study error** is error introduced when the differences between the target and study pop-

ulations affect the applicability of the study to the target population

Formulas

You're probably here to learn mathematics, so you may be eager to start working with some numbers and formulas. For a given population of people, we'll define these variables:

- Let TP be the number of true positives in the population.
- Let TN be the number of true negatives in the population.
- Let FP be the number of false positives in the population.
- Let FN be the number of false negatives in the population.

The sensitivity or true positive rate (TPR) is

$$\text{sensitivity (aka } TPR) = \frac{TP}{\text{everyone who actually has the disease}} = \frac{TP}{TP + FN}.$$

Also, the specificity or the true negative rate (TNR) is

$$\text{specificity (aka } TNR) = \frac{TN}{\text{everyone who doesn't have the disease}} = \frac{TN}{FP + TN}.$$

These names come from how well the test can correctly classify individuals who really have the condition and how well they can correctly classify individuals who really don't have the condition.

Note: The sensitivity and specificity of a diagnostic test is generally considered to be stable for a given test, and an inherent property of the test.

Warning: The true positive rate is not the fraction of true positives out of all positives. The true negative rate is not the fraction of true negatives out of all negatives. What is the denominator in each case instead?

Example 4

(a) A test has a specificity of 16%. Out of 600 people who do not have the disease, how many people have a false positive?

(b) A test has a sensitivity of 98%. Out of 800 people who have the disease, how many people have a false negative?

(c) A test has a specificity of 75%. If 572 people do not have the disease, how many of them have a false positive?

Solution:

(a) Based on the specificity, 16% of people without the disease will receive a true negative result. The rest of the people without the disease will have a false positive result. The percentage of people who will have a false positive result is therefore $(1 - \text{specificity})$. This works out to be $100\% - 16\% = 84\%$. So

$$\begin{aligned}\#\text{ false positives} &= (1 - \text{specificity}) \times 600 \\ &= (1 - 0.16) \times 600 \\ &= 0.84 \times 600 \\ &= 504.\end{aligned}$$

There are 504 false positives.

(b) Of all the people with the disease, some will be true positives and the others will be false negatives. The percentage of people with the disease who are false negatives is $(1 - \text{specificity})$. This works out to be $100\% - 98\% = 2\%$. So

$$\begin{aligned}\#\text{ false negatives} &= (1 - \text{sensitivity}) \times 800 \\ &= (1 - 0.02) \times 800 \\ &= 0.98 \times 800 \\ &= 784\end{aligned}$$

There are 784 false negatives.

(c) Of all the people without the disease, some will be true negatives and the others will be false positives. The percentage of people without the disease who are false positives is

(1 − specificity). This works out to be 100% − 75% = 25%. So

$$\begin{aligned}
\#\text{ false positives} &= (1 - \text{specificity}) \times 572 \\
&= (1 - 0.75) \times 572 \\
&= 0.25 \times 572 \\
&= 143
\end{aligned}$$

There are 143 false positives.

Example 5

Suppose that a test for a certain condition has a sensitivity of 82% and a specificity of 99%.

1. What is the probability that a person who actually has the disease will be correctly identified by the test?
   (a) 100%
   (b) 99%
   (c) 82%
   (d) Not enough information
2. Suppose that someone receives a positive test result. What is the probability that this person actually has the disease?
   (a) 100%
   (b) 99%
   (c) 82%
   (d) Not enough information
3. Suppose that someone receives a negative test result. What is the probability that this person doesn't have the disease?
   (a) 100%
   (b) 99%
   (c) 82%
   (d) Not enough information

Solution:

1. By definition, the probability that a person who actually has the disease will be correctly identified by the test is the test's sensitivity. So this answer is (a), 82%.
2. Just knowing the sensitivity and specificity of the test is insufficient in calculating the probability that someone who has a positive test result actually has the disease. (Take a look at the warning on page 6.) This probability is called the positive predictive value, which is introduced later in the lesson.
3. Just knowing the sensitivity and specificity of the test is insufficient in calculating the probability that someone who has a negative test result actually has the disease. (Take a look at the warning on page 6.) This probability is called the negative predictive value, which is not a part of this lesson.

Example 6

Circle which property of a diagnostic test each statement provides. Then state the value of that property.

(a) Out of 100 healthy kids tested, 2 have a false positive strep test.

Sensitivity   Specificity

Value: ____________________

(b) Two-thirds of patients with atherosclerosis will have a positive exercise stress test.

Sensitivity   Specificity

Value: ____________________

Solution:

1. Since the 100 kids that are tested are healthy, they can either receive a false positive or a true negative test result. If 2 out of every 100 healthy kids has a false positive strep test, then 98 out of every healthy kids have a true negative strep test. $\boxed{\text{Specificity}}$ is the appropriate term here: the specificity of the test is $\frac{TN}{TN+FP} = \frac{98}{100} = 98\%$.
2. Patients with atherosclerosis can either receive a true positive or a false negative result when an exercise stress test is performed. Since $\frac{2}{3}$ of patients with atherosclerosis will have

a positive exercise stress test, then it can be implied that $\frac{1}{3}$ of patients with atherosclerosis will have a negative exercise stress test. $\boxed{\text{Sensitivity}}$ is the appropriate term here: the sensitivity of the test is $\frac{TP}{TP+FN} = \frac{\frac{2}{3}}{1} = \frac{2}{3}$. (Here we can consider TP and FN as fractions of the population, whose values are $\frac{2}{3}$ and $\frac{1}{3}$ respectively.)

The Importance of Sensitivity and Specificity

Stop and Think. Why might it be important for people who don't have a disease to know that they don't have it? Why might it be important for people who have a disease to know that they have it?

Sample answer: It can be important for people who don't have a disease to know that they don't have it because a false positive can lead to

- Costly and risky treatment or further testing
- Negative psychological impacts (e.g, from social stigma associated with the condition)

It can be important for people who have a disease to know that they have it because a false negative can lead to

- Serious delays in proper diagnosis and treatment of the disease
- Possibly place others at risk in the case of a communicable disease

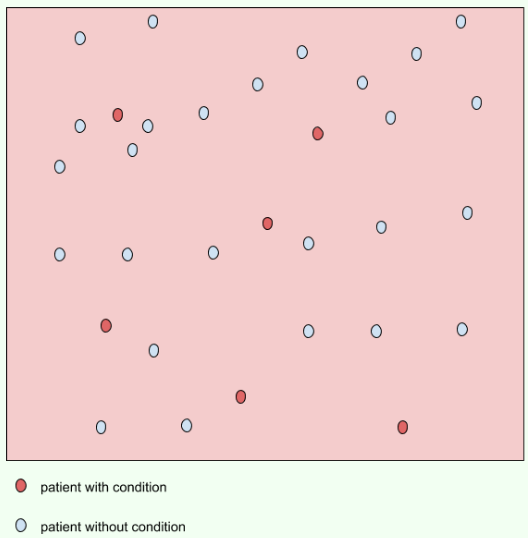
As a specific example, consider the SARS-CoV-2 rapid antigen test. During the COVID-19 pandemic, if someone has a positive rapid antigen test result but doesn't actually have COVID-19 (i.e., they have a **false positive** test result), they may be asked to quarantine, which would affect their quality of life and could result in lost productivity in the workplace.

On the other hand, if someone has a negative rapid antigen test result but actually has COVID-19 (i.e., they have a **false negative** test result), they may develop a false sense of security and engage in risky behavior such as attending large gatherings, which puts others at risk of infection.

We would like now to develop the skill of interpreting diagrams to calculate the sensitivity and specificity of a test.

Example 7

A group of people are tested and the test results are represented in the following figure. A red dot indicates that the patient has the condition. A white dot indicates that the patient does not actually have the condition but has tested positive. The red background indicates the area where the test produces a positive test result for the patient. What is the sensitivity and specificity of the test based on these results?[1]



Solution: In this case we have

- There are 25 **false positives** (white dots).
- There are 6 **true positives** (red dots).
- There are 0 **false negatives** (since the test predicts everyone has the condition).
- There are 0 **true negatives** (since the test predicts everyone has the condition).

Therefore, the sensitivity is

$$TPR = \frac{TP}{TP + FN} = \frac{6}{6 + 0} = 100\%.$$

Also, the specificity is

$$TNR = \frac{TN}{TN + FP} = \frac{0}{0 + 25} = 0\%.$$

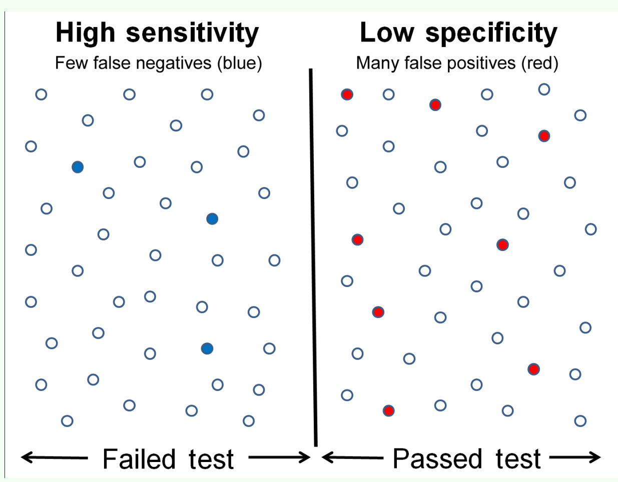
The previous example shows that we cannot determine how "good" a test is using sensitivity by itself. A test with a very high sensitivity (e.g., 100%) can be completely useless in determining who actually has the condition if it has a low specificity. One example of this is a hypothetical test that unilaterally produces a positive result for everyone (like in Example 7).

High sensitivity just means that most people with the disease are correctly identified as having the disease by the test. A test with 100% sensitivity means that *everyone* who has the disease is correctly identified by the test.

High specificity just means is that most people without the disease are correctly identified as not having that disease by the test. A test with 100% specificity means that *everyone* who does not have the disease is correctly identified by the test.

Example 8

A group of people are tested for a disease and the results are represented in the following figure. To the left of the middle line, each dot represents an individual who has failed the test. False negatives are coloured in blue and true negatives are not coloured. To the right of the middle line, each dot represents an individual who has passed the test. False positives are coloured in red and true positives are coloured in white. What is the sensitivity and specificity of the test based on these results?[2]



Solution: In this case we have

- There are 3 false negatives (blue dots on the left).

[1]**Image source:** Kchusap, CC BY-SA 4.0, via Wikimedia Commons

- There are 37 true negatives (uncoloured dots on the left).
- There are 8 false positives (red dots on the right).
- There are 32 true positives (uncoloured dots on the right).

Therefore, the sensitivity is

$$TPR = \frac{TP}{TP + FN} = \frac{32}{32 + 3} = \frac{32}{35} \approx 91.4\%.$$

Also, the specificity is

$$TNR = \frac{TN}{TN + FP} = \frac{37}{37 + 8} = \frac{37}{45} \approx 82.2\%.$$

Sometimes we need to make use of the fact that the number of true positives and false positives add up to the total number of positives and that the number of true negatives and false negatives add up to the total number of negatives.

Positive Predictive Value

Specificity and sensitivity are great—ideally we want almost everyone who doesn't have the disease to be correctly identified (high specificity) and almost everyone who does have the disease to be correctly identified (high sensitivity).

But what if you actually had a positive test? Should you be worried? If you had a positive test result, what's the probability that you actually had the disease?

According to one report[3], the chance that someone actually has breast cancer given that they have a positive screening mammography test is less than 15%, which means that a positive test result is more likely to indicate *not* having breast cancer than to indicate having it. This probability increases with age, partly because breast cancer is more common in older people. As we shall see, how common a disease is in a population affects the chance that a positive test result correctly indicates disease.

The chance that someone who has received a positive test result actually has the disease is called the **positive predictive value**. In the breast cancer example, the screening mammography has a positive predictive value of less than 15%.

[2] **Image source**: Rmostell, CC0, via Wikimedia Commons

[3] Retrieved from the Canadian Partnership Against Cancer website here.

A diagnostic test provides information that gives us more evidence to support the hypothesis that the person has the disease. For example, if the probability that an asymptomatic 20-year-old male in Toronto during March 2021 has COVID-19 is 2%, then the probability that an asymptomatic 20-year-old male in Toronto during 2021 who's received a *positive result on a rapid antigen test* has COVID-19 could be 30%[4], which would be the positive predictive value for that population. If the test is any good, then the positive predictive value would be higher than the probability that someone has the disease without any test information. Updating the probability (here, probability is interpreted as a degree of belief) that something is true (e.g., someone has a disease) using information such as that from a diagnostic test is known as **Bayesian inference**.

Notice that 30% is still less than a 50-50 chance. It would still be more likely for someone who fits that description to be free of COVID, even if they have a positive COVID-19 rapid antigen test result. This is due to the low **prevalence** of COVID-19. Prevalence is the percentage of everyone who actually has the disease. Recall that TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives, and TN is the number of true negatives. The formula for prevalence looks like

$$\text{Prevalence} = \frac{TP + FN}{TP + FP + TN + FN} = \%\text{ of the population with the disease.}$$

The prevalence represents our initial degree of belief that someone has a disease. Bayesian inference is used to update this degree of belief (increase it or decrease it) depending on new information gained from a test result. To fail to consider the initial degree of belief, the disease prevalence, when interpreting test results is known as **the base rate fallacy**, **base rate neglect**, or the **false positive paradox**. Administrators of diagnostic tests may be especially likely to commit this fallacy when they change from testing in a high-prevalence environment to a low-prevalence environment (or vice versa) or when they otherwise do not have an accurate sense of the prevailing prevalence of the disease.

The important point is that the probability (or degree of belief) that someone has a disease is updated by the diagnostic test, but is also heavily influenced by the starting point, which is the prevalence of the disease in the first place.

Of course, if we already know how many true and false positives there are, we can just calculate

[4] **Source**: The ObG Project

positive predictive value directly:

$$\text{Positive predictive value} = \frac{TP}{TP + FP}$$

If we don't know the exact number of true positives and false positives, we can calculate the positive predictive value using percentages

- $TP = \text{sensitivity} \times \text{prevalence}$

  The percentage of everyone who is a true positive is the percentage of everyone who has the disease multiplied by the sensitivity (i.e., the percentage of everyone who has the disease that has a positive test result).

- $FP = (1 - \text{specificity}) \times (1 - \text{prevalence})$

  The percentage of everyone who is a false positive is the percentage of everyone who does not have the disease multiplied by $(1 - \textit{sensitivity})$ *(which is the percentage of everyone without the disease who does* ***not*** *have a true negative result and therefore has a false positive result)*

Replacing TP and FP with these expressions, we can write the positive predictive value in terms of sensitivity, specificity, and prevalence:

$$\text{Positive predictive value} = \frac{TP}{TP + FP} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

This derivation of this formula is an application of **Bayes' Theorem**.

So it turns out that specificity and sensitivity is not enough information to determine the positive predictive value; we need to know the prevalence of the disease as well.

Stop and Think. Would specificity (catching people *without* the disease) or sensitivity (catching people *with* the disease) have a greater effect on the positive predictive value of a test?

Due to the relatively low prevalence of most diseases and conditions, specificity matters more than sensitivity in affecting the positive predictive value—that is, in order to increase the positive predictive value of a test, it is more helpful to have a specific test that identifies most people without the disease correctly than it is to have a sensitive test.

Notice that the sensitivity will increase the number of true positives (TPs), which appear in both the numerator and the denominator. Since the % increase in the numerator is greater than the % increase in the denominator, which has a term added to the TPs, then this will increase the PPV overall. Most of the time, however, the denominator is dominated by the high number of FPs, which is represented by the $(1 - \text{specificity}) \times (1 - \text{prevalence})$ term, due to the low prevalence of the disease. In other words, there are usually much more false positives than true positives, and so decreasing the number of false positives relative to true negatives (i.e., raising specificity) often has a bigger effect on the PPV than increasing the number of true positives relative to false negatives (i.e., raising sensitivity).

Example 9

(a) Calculate the prevalence of a disease if, in a population of 10000 people, 50 people have the disease.

(b) If in the same population of 10000 people, 250 people have the disease, then what is the prevalence of the disease?

(c) Suppose that the prevalence of a disease in a population of 2400 people was 32%. How many of the 2400 people would we expect to have the disease?

Solution:

(a) The prevalence of the disease is

$$\begin{aligned}\text{Prevalence} &= \frac{\#\text{ people with the disease}}{\#\text{ people in the population}}\\ &= \frac{50}{10000}\\ &= 0.005\\ &= 0.5\%.\end{aligned}$$

It's usually best to leave the prevalence as a percentage as we have done above.

(b) The prevalence of the disease is

$$\begin{aligned}\text{Prevalence} &= \frac{\#\text{ people with the disease}}{\#\text{ people in the population}}\\ &= \frac{250}{10000}\\ &= 0.025\\ &= 2.5\%.\end{aligned}$$

(c) We expect 32% of the people to have the disease. Thus, we have

$$\begin{aligned}\text{\# people with the disease} &= \text{Prevalence} \times (\text{\# people in the population}) \\ &= 32\% \times 2400 \\ &= 768.\end{aligned}$$

So we expect that 768 of the 2400 people have the disease.

Example 10

If 13 out of 15 positives are true positives and 16 out of 17 negatives are true negatives. What is the prevalence of the disease based on the test data?

Solution: To calculate the prevalence of the disease, let's first determine the number of true negatives, false negatives, true positives, and false positives:

- There are 13 true positives
- The other $15 - 13 = 2$ positives must be false positives. So there are 2 false positives.
- There are 16 true negatives.
- The other $17 - 16 = 1$ negatives must be false negatives. So there is 1 false negative.

Stop and Think. Do you remember which one of the above test results means that the person actually has the disease? If a person actually has a disease, they can either have a true positive or a false negative test result. So there are $13 + 1 = 14$ people with the disease based on the above numbers.

There are a total number of $15 + 17 = 32$ people tested, and so the prevalence is

$$\begin{aligned}\text{Prevalence} &= \frac{\text{\# people with the disease}}{\text{\# people in the population}} \\ &= \frac{14}{32} \\ &= 43.75\%\end{aligned}$$

Example 11

Suppose that a test for a certain condition has a sensitivity of 100% and a specificity of 95%.

(a) Out of 1000 people tested, 80 really have the disease and the rest don't. What is the prevalence of the disease and the expected positive predictive value? Express your answer as a percentage rounded to one decimal place.

(b) Out of 1000 people tested, 20 really have the disease and the rest don't. What is the prevalence of the disease and the expected positive predictive value? Express your answer as a percentage rounded to one decimal place.

(c) Take a look at your answers for parts (a) and (b). Is a higher prevalence associated with a higher positive predictive value? Or was it associated with a lower positive predictive value?

Solution:

1. The prevalence of the disease is

$$\frac{\#\text{ of people who have the disease}}{\#\text{ people in the population}} = \frac{50}{1000} = 8\%.$$

To find the positive predictive value, we first need to determine the number of true positives and the number of false positives. Since the sensitivity of the test is 100%, then the number of true positives is 80 (everyone who has the disease tests positive). Since the specificity of the test is 95%, the the number of false positives is $(100\% - 95\%) \times 920 = 46$, where 920 is the number of people who don't have the disease. Thus, the positive predictive value is

$$\text{Positive predictive value} = \frac{TP}{TP + FP} = \frac{50}{50 + 46} = \frac{50}{96} \approx 52.1\%.$$

2. The prevalence of the disease is

$$\frac{\#\text{ of people who have the disease}}{\#\text{ of people in the population}} = \frac{20}{1000} = 2\%$$

To find the positive predictive value, we first need to determine the number of true positives and the number of false positives. Since the sensitivity of the test is 100%, then the number of true positives is 20 (everyone who has the disease tests positive). Since the specificity of the test is 95%, the number of false positives is $(100\% - 95\%) \times 980 = 49$, where 49 is the number of people who don't have the disease. Thus, the positive predictive value is

$$\text{Positive predictive value} = \frac{TP}{TP + FP} = \frac{20}{20 + 49} = \frac{20}{69} \approx 29.0\%$$

3. The positive predictive value goes down substantially as the prevalence decreases from 8% to 2%.

Example 12

A test for hypertension has a specificity of 90% and a sensitivity of 90%.

(a) What is the probability that a person with a positive test result actually has the disease if the prevalence of hypertension is 20%?

(b) What is the probability that a person with a positive test result actually has the disease if the prevalence of hypertension is 1%?

(c) Take a look at your answers for parts (a) and (b). Is a higher prevalence associated with a higher positive predictive value? Or was it associated with a lower expected positive predictive value?

Solution:

(a) If the prevalence of hypertension is 20%, the probability that someone with a positive test result actually has the disease is

$$\begin{aligned}\text{Positive predictive value} &= \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}\\ &= \frac{0.90 \times 0.20}{0.90 \times 0.20 + (1 - 0.90) \times (1 - 0.20)}\\ &= \frac{0.18}{0.18 + 0.1 \times 0.8}\\ &= \frac{0.18}{0.18 + 0.08}\\ &= 69.2\%\end{aligned}$$

(b) If the prevalence of hypertension is 1%, the probability that someone with a positive test result actually has the disease is

$$\begin{aligned}
\text{Positive predictive value} &= \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})} \\
&= \frac{0.90 \times 0.01}{0.90 \times 0.01 + (1 - 0.90) \times (1 - 0.01)} \\
&= \frac{0.009}{0.009 + 0.1 \times 0.99} \\
&= \frac{0.009}{0.009 + 0.099} \\
&= 8.3\%.
\end{aligned}$$

(c) The positive predictive value goes down substantially as the prevalence decreases from 20% to 1%.

Counterintuitive Results

False positives can be more probable than true positives (i.e., the positive predictive value is low) when the prevalence of disease is low. This is because there may be way more people who don't have the disease than people who do. In this case, the test will produce many false positives unless its specificity is very high. This counter-intuitive result is called the **false-positive paradox**.

Note that COVID-19 rapid antigen tests tend to have the opposite problem. These tests often have a very high specificity (around 99%) but a relatively low sensitivity such as (70%). Depending on disease prevalence, this means that many negative test results may actually have a relatively high probability of being a false negative (i.e., the person actually has COVID-19), and the positive predictive value of the test is very high (80% to 100%), especially among symptomatic patients[a].

[a]**Source:** The ObG Project

Suppose that you are given the following problem.

A test for a certain condition has a sensitivity of 82% and a specificity of 90% in a given population. Out of 100 people tested, 20 test positive and the rest test negative. What is the expected prevalence of the disease?

This is actually a very hard problem to solve!

Just because 20 people tested test positive doesn't mean that they are all true positives. Also, there may be some false negatives among the 80 negative test results (i.e., some of the people who have received a negative test result actually have the condition). In fact, we know this to be the case since neither the sensitivity nor the specificity of the test is 100%.

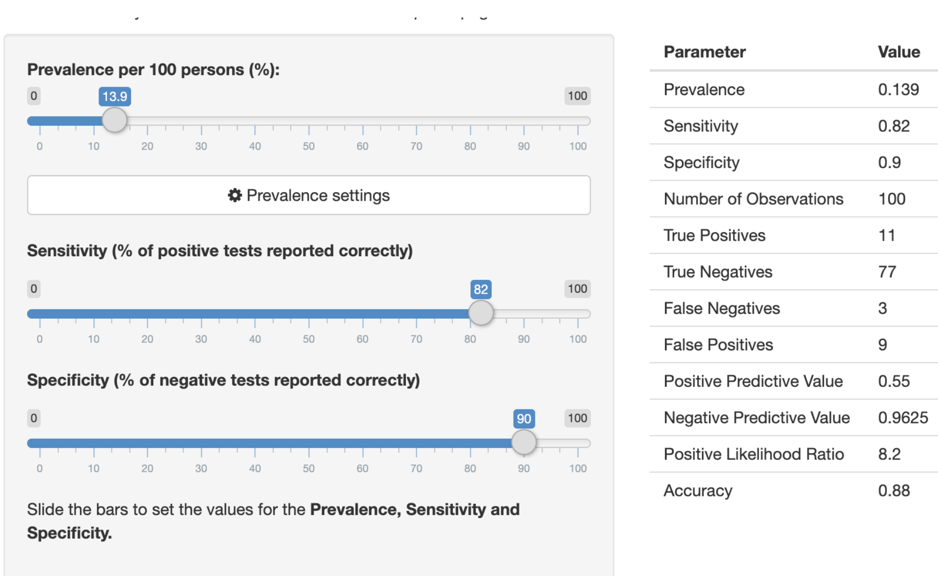
We will leave computing an exact answer algebraically as an exercise. However, we can use an online simulator to see the interplay between sensitivity, specificity, and prevalence. This online simulator also shows us what the positive predictive value is as well as the exact number of false positives, true positives, false negatives, and true negatives.

The online simulator provides three sliders: one each for prevalence, sensitivity, and specificity. Let's set the sensitivity and specificity to that specified in the problem.

Notice the two curves on the right-hand side. The blue curve is the positive predictive value, and it goes from 0% when the prevalence is 0% all the way up to 100% when the prevalence is 100%. This highlights the point that **positive predictive value depends on prevalence!** Notice that the blue curve rises really fast between a prevalence of 0% and 25% and then goes up slower as prevalence continues to increase. This shows that the positive predictive value is relatively high when the prevalence reaches around 25%, but between a prevalence of 0% and 25% the positive predictive value can be quite low and is affected greatly even by a small change in prevalence.

As we move the third slider, to control the prevalence, the red dot on the blue line changes. This shows exactly the point that our prevalence corresponds to on the blue curve for positive predictive value.

The simulator shows the exact number of true positives, true negatives, false positives, and false negatives out of the 100 observations. (The number of observations can be changed by going into "Prevalence settings"). We can continue to adjust the prevalence slider until the true positives and the false positives add up to 20 (which is what is stated in the problem). This occurs when the prevalence is approximately 13.9%. When the prevalence is about 13.9%, there are 11 true positives and 9 false positives, giving a positive predictive value of only 55%. So just over half of the 20 positive tests are true positives!



Parameter	Value
Prevalence	0.139
Sensitivity	0.82
Specificity	0.9
Number of Observations	100
True Positives	11
True Negatives	77
False Negatives	3
False Positives	9
Positive Predictive Value	0.55
Negative Predictive Value	0.9625
Positive Likelihood Ratio	8.2
Accuracy	0.88

Figure 1: Screenshot of the three sliders for sensitivity, specificity, and prevalence as well as the listing of parameters in the simulator

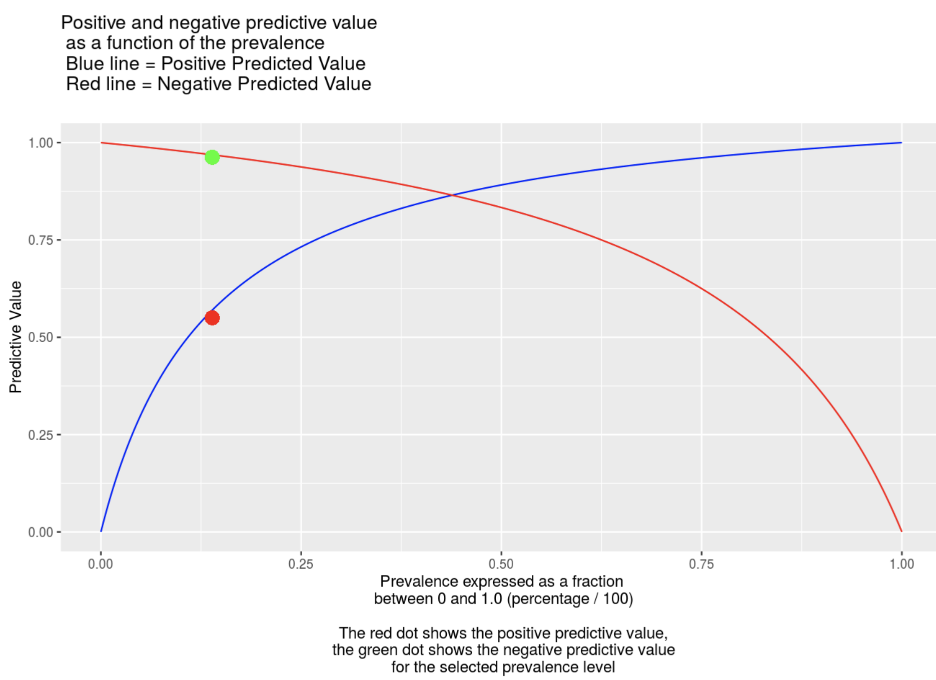


Figure 2: Screenshot of the graph of positive (and negative) predictive value versus prevalence in the simulator.