Positive Predictive Value (Bayes' Theorem)

Chapter 2, Lab 3: Solutions

OpenIntro Biostatistics

Topics

- Positive Predictive Value
- Bayes' Theorem
- Simulation

The **positive predictive value (PPV)** of a diagnostic test is the probability that a person has a disease, given that they tested positive for it. This lab illustrates three common approaches for calculating PPV: contingency tables, tree diagrams, and simulation. The last two questions are more concept-oriented and probe the relationships between prevalence, sensitivity, specificity, PPV, and NPV.

The material in this lab corresponds to Section 2.2.5 of *OpenIntro Biostatistics*.

Introduction

Suppose a child tests positive for trisomy 21 from a cell-free fetal DNA (cfDNA) test. What is the probability that the child does have trisomy 21, given the positive test result?

This conditional probability, $P(D|T^+)$, where D is the event of having the disease and T^+ is the event of a positive test, has a specific name: **positive predictive value (PPV)** . Similarly, the **negative predictive value (NPV)**, $P(D^c|T^-)$, is the probability that disease is absent when test results are negative. These values are of particular importance to clinicians in assessing patient health.

The characteristics of diagnostic tests are given with the reverse conditional probabilities—the probability that the test correctly returns a positive test in the presence of disease and a negative result in the absence of disease. The probability $P(T^+|D)$ is referred to as the **sensitivity** of the test and $P(T^-|D^c)$ is the **specificity** of the test.[1] Given these two values and the marginal probability of disease $P(D)$, also known as the **prevalence**, it is possible to calculate both the PPV and NPV of a diagnostic test.

- Sensitivity: Of 1,000 children with trisomy 21, approximately 980 test positive.

- Specificity: Of 1,000 children without trisomy 21, approximately 995 test negative.

- Prevalence: Trisomy 21 occurs with a rate of approximately 1 in 800 births.

There are several possible strategies for approaching this type of calculation: 1) creating a contingency table for a large, hypothetical population, 2) running a simulation, 3) using an algebraic approach with Bayes' Theorem.

---

[1]Note that the sensitivity can also be referred to as the probability of a true positive and the specificity is the probability of a true negative. Thus, the probability of a false negative equals (1 - sensitivity) and the probability of a false positive equals (1 - specificity).

1. **Table-based approach**. Constructing a contingency table for a large, hypothetical population offers an intuitive way to understand the distribution of disease incidence and test outcome. Suppose that the total population is 100,000. The following table can be filled in based on the known information about disease prevalence, test specificity, and test sensitivity.

	Disease Present	Disease Absent	Sum
Test Positive	122.5	499.375	621.875
Test Negative	2.5	99,375.62	99,378.12
Sum	125	99,875	100,000

   a) Calculate the two column totals—out of 100,000 children, how many are expected to have trisomy 21? How many are expected to not have trisomy 21?

If trisomy 21 occurs in 1 out of 800 births, then $(1/800)(100,000) = 125$ children are expected to have trisomy 21 in a population of 100,000. It then follows that $100,000 - 125 = 99,875$ children are expected to not have trisomy 21; this could also be calculated by $(799/800)(100,000) = 99,875$.

   b) The most efficient way to compute each table value is to use R as a calculator. Run the following code chunks and populate the rest of the table cells. Note that the first chunk simply re-calculates the values from part a).

```
#parameters
prevalence = 1/800
sensitivity = 0.980
specificity = 0.995
population.size = 100000

#expected number with trisomy 21
expected.cases = population.size * prevalence
expected.cases
```

```
## [1] 125
```

```
#expected number without trisomy 21
expected.noncases = population.size - expected.cases
expected.noncases
```

```
## [1] 99875
```

```
#expected number with trisomy 21, tested positive (true pos)
expected.true.positives = expected.cases * sensitivity
expected.true.positives
```

```
## [1] 122.5
```

```r
#expected number without trisomy 21, tested positive (false pos)
expected.false.positives = expected.noncases * (1 - specificity)
expected.false.positives
```

```
## [1] 499.375
```

```r
#total expected positives
total.expected.positives = expected.true.positives + expected.false.positives
total.expected.positives
```

```
## [1] 621.875
```

```r
#expected number with trisomy 21, tested negative (false neg)
expected.false.negatives = expected.cases * (1 - sensitivity)
expected.false.negatives
```

```
## [1] 2.5
```

```r
#expected number without trisomy 21, tested negative (true neg)
expected.true.negatives = expected.noncases * specificity
expected.true.negatives
```

```
## [1] 99375.62
```

```r
#total expected negatives
total.expected.negatives = expected.true.negatives + expected.false.negatives
total.expected.negatives
```

```
## [1] 99378.12
```

c) Using values from the table, calculate an estimate of the probability that a child who tests positive actually has trisomy 21. *Hint:* Think about how the definition of conditional probability could be applied here... $P(D|T^+) = \frac{P(D \text{ and } T^+)}{P(T^+)}$

$$P(D|T^+) = \frac{P(D \text{ and } T^+)}{P(T^+)} = \frac{\frac{122.5}{100,000}}{\frac{621.875}{100,000}} = \frac{122.5}{621.875} = 0.197$$

The quantity $P(D|T^+)$ can be estimated as the expected number of true positive tests divided by the total expected number of positive tests.

```r
#ppv
ppv = expected.true.positives/total.expected.positives
ppv
```

```
## [1] 0.1969849
```

2. **Simulation approach**. Rather than building a population based on expected values, a population can be simulated using the provided probabilities. This approach was introduced in the previous lab.

The following code creates a simulated dataset of 100,000 individuals that each have a disease status and test result. Run the code then answer the following questions. There are comments in the code that correspond to the following questions; i.e., when answering part a), look for the comment (part a).

```
#define parameters
population.size = 100000
prevalence = 1/800
sensitivity = 0.980
specificity = 0.995

#create empty vectors to store results
disease.status = vector("numeric", population.size)
test.result = vector("numeric", population.size)

#set the seed for a pseudo-random sample
set.seed(2018)

#assign disease status (part a)
disease.status = sample(c(0,1), size = population.size,
                    prob = c(1 - prevalence, prevalence),
                    replace = TRUE)

#assign test result (part b)
for(k in 1:population.size){

  if(disease.status[k] == 0){
test.result[k] = sample(c(0,1), size = 1,
                    prob = c(specificity, 1 - specificity))
  }

  if(disease.status[k] == 1){
test.result[k] = sample(c(0,1), size = 1,
                    prob = c(1 - sensitivity, sensitivity))
  }

}

#create matrix of disease status and test result (part c)
disease.status.and.test.result = cbind(disease.status, test.result)

#create a table of test result by disease status
addmargins(table(test.result, disease.status))

##             disease.status
## test.result      0      1    Sum
##           0   99405      3  99408
##           1     466    126    592
```

4

```
##         Sum  99871    129 100000
```

```
#calculate ppv (part d)
ppv = sum(test.result[disease.status == 1])/sum(test.result)
ppv
```

```
## [1] 0.2128378
```

```
#calculate npv (part e)
npv = ((population.size - sum(disease.status)) - sum(test.result[disease.status
                                                   == 0])) /
  (population.size - sum(test.result))
npv
```

```
## [1] 0.9999698
```

a) Explain how `sample()` is being used to fill in `disease.status`. If an individual is assigned a 0, what is their disease status?

Disease status is assigned based on prevalence. If an individual is assigned a 0, they do not have the disease; if they are assigned a 1, they have the disease. The value 0 is sampled with probability (1 - prevalence), while 1 is sampled with probability (prevalence).

b) How is test outcome assigned if an individual has disease status 0? How is test outcome designed if an individual has disease status 1?

If an individual has the disease, test outcome is assigned based on sensitivity and the false negative rate, (1 - specificity). If an individual does not have the disease, test outcome is assigned based on specificity and the false positive rate, (1 - sensitivity). In both cases, 0 corresponds to a negative test result and 1 to a positive test result.

c) Take a look at `disease.status.and.test.result`. What does a single row ith a 0 in both columns represent?

A single row with a 0 in both columns represents an individual who does not have trisomy 21 and tested negative; i.e., a true negative result.

d) Explain the line of code used to calculate PPV based on the results of the simulation.

The numerator consists of the total number of individuals who tested positive and have the disease; `sum(test.result[disease.status == 1])` calculates the sum of `test.result` only for the indices where `disease.status` equals 1. The denominator consists of all individuals who tested positive.

e) Is the PPV calculated through the simulation different from the result using the table method? If so, explain why the two numbers differ.

The PPV calculated through this simulation method is slightly different than the result from the table method; this method returned a PPV of 0.213 versus a value of 0.197. Some variation is to be expected with the simulation; if the simulation were repeated with a larger population size, then the returned result is likely to be closer to 0.197.

f) Estimate the NPV based on the results of the simulation.

3. **Algebraic approach**. The expanded formula for PPV is based upon the definition of conditional probability, and rewriting joint probabilities as the product of a conditional and marginal probability by applying the general multiplication rule. The logic behind the formula can be illustrated via a tree diagram.

$$P(D|T^+) = \frac{P(D \text{ and } T^+)}{P(T^+)}$$

$$= \frac{P(D \text{ and } T^+)}{P(D \text{ and } T^+) + P(D^c \text{ and } T^+)}$$

$$= \frac{P(T^+|D)P(D)}{[P(T^+|D) \times P(D)] + [P(T^+|D^c) \times P(D^c)]}$$

a) Draw a tree diagram that organizes the four possible combinations of disease incidence and test outcome: D and T^+, D and T^-, D^C and T^+, D^C and T^-.
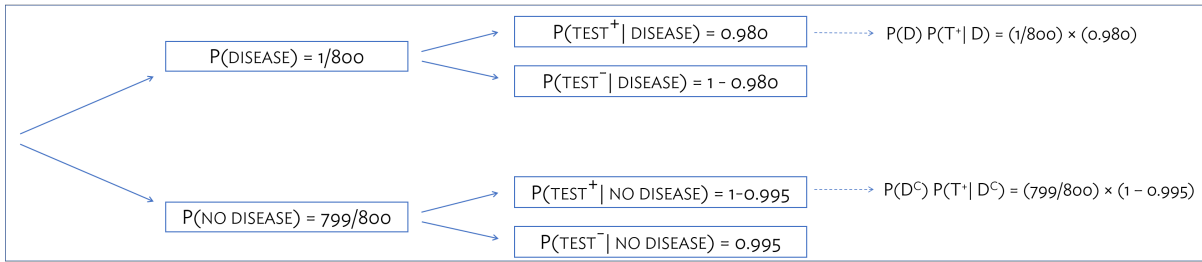
b) Use the formula to calculate $P(D|T^+)$.

$$P(D|T^+) = \frac{P(D \text{ and } T^+)}{P(T^+)}$$

$$= \frac{P(D)P(T^+|D)}{P(T^+)}$$

$$= \frac{P(D)P(T^+|D)}{P(D)P(T^+|D) + P(D^C)P(T^+|D^C)}$$

$$= \frac{(1/800)(0.980)}{(1/800)(0.980) + (799/800)(1 - 0.995)}$$

$$= 0.197$$

The topmost branch of the tree illustrates how $P(D \text{ and } T^+)$ can be calculated by using the general multiplication rule and multiplying $P(D)$ by $P(T^+|D)$. There are only two "ways" to test positive–either as a true positive or as a false positive, and these are disjoint events; thus, $P(T^+) = P(D \text{ and } T^+) + P(D^C \text{ and } T^+)$.

c) The formula can be generalized in terms of events A and B, where

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}.$$

In this form, it is commonly known as Bayes' Theorem. Apply Bayes' Theorem to calculate the NPV.

$$
\begin{aligned}
P(D^c|T^-) &= \frac{P(D^c \text{ and } T^-)}{P(T^-)} \\
&= \frac{P(D^c)P(T^-|D)^c}{P(T^-)} \\
&= \frac{P(D^c)P(T^-|D^c)}{P(D^c)P(T^-|D^c) + P(D)P(T^-|D)} \\
&= \frac{(799/800)(0.995)}{(799/800)(0.995) + (1/800)(1 - 0.980}} \\
&= 0.999974
\end{aligned}
$$

Challenge Questions

4. The strongest risk factor for breast cancer is age; as a woman gets older, her risk of developing breast cancer increases. The following table shows the average percentage of American women in each age group who develop breast cancer, according to statistics from the National Cancer Institute. For example, approximately 3.56% of women in their 60's get breast cancer.

Table 1: Prevalence of Breast Cancer by Age Group

Age Group	Prevalence
30 - 40	0.0044
40 - 50	0.0147
50 - 60	0.0238
60 - 70	0.0356
70 - 80	0.0382

A mammogram typically identifies a breast cancer about 85% of the time, and is correct 95% of the time when a woman does not have breast cancer.

a) If a woman in her 60's has a positive mammogram, what is the likelihood that she has breast cancer? Solve this problem algebraically.

Use Bayes' rule to solve for $P(D|T^+)$.

$$
\begin{aligned}
P(D|T^+) &= \frac{P(T^+|D) \cdot P(D)}{[P(T^+|D) \cdot P(D)] + [P(T^+|D^C) \cdot P(D^C)]} \\
P(\text{cancer}|\text{mammogram}^+) &= \frac{P(\text{mammogram}^+| \text{cancer}) \cdot P(\text{cancer})}{[P(\text{mammogram}^+| \text{cancer})] + [P(\text{mammogram}^+| \text{no cancer})]} \\
&= \frac{0.85 \cdot (0.0356)}{0.85 \cdot (0.0356) + (1 - 0.95) \cdot (1 - 0.0356)} \\
&= 0.3856
\end{aligned}
$$

b) Use an R simulation to simulate the results for administering mammograms to a population of 100,000 women in their 30's. How many women in this hypothetical population are expected to test positive for breast cancer? Estimate the PPV of a mammogram for a woman in her 30's.

This simulation results in 5,426 positive tests and an estimated PPV of 0.0735.

```r
#define parameters
population.size = 100000
prevalence = 0.0044
sensitivity = 0.85
specificity = 0.95

#create empty vectors to store results
disease.status = vector("numeric", population.size)
test.result = vector("numeric", population.size)

#set the seed for a pseudo-random sample
set.seed(2018)

#assign disease status
disease.status = sample(c(0,1), size = population.size,
                   prob = c(1 - prevalence, prevalence),
                   replace = TRUE)

#assign test result
for(k in 1:population.size){

  if(disease.status[k] == 0){
test.result[k] = sample(c(0,1), size = 1,
                     prob = c(specificity, 1 - specificity))
  }

  if(disease.status[k] == 1){
test.result[k] = sample(c(0,1), size = 1,
                     prob = c(1 - sensitivity, sensitivity))
  }

}

#calculate expected number of positive tests
sum(test.result)
```

```
## [1] 5426
```

```r
#calculate ppv
ppv = sum(test.result[disease.status == 1])/sum(test.result)
ppv
```

```
## [1] 0.07353483
```

c) Using whatever methods you wish, calculate the PPV for each age group; show your work. Describe any trends you see in the PPV values as prevalence changes. Explain the reason for the trends in language that someone who has not taken a statistics course would understand.

As the prevalence of breast cancer increases, PPV also increases. If more women have the disease, the chance of a positive test result being a true positive increases (and the chances of the result being a false positive decreases).

To explain the reason for this trend, it is helpful to refer to the PPV formula and consider extreme cases. Suppose that the prevalence of a disease is very low, and extremely close to 0; the numerator would also be close to 0, causing PPV to be low. In the other direction, if prevalence is very high (almost 1), then the second term in the denominator that contains *(1 - prevalence)* essentially disappears, and the numerator divided by the denominator will be very high (almost 1).

Another way to think about this is to focus on the denominator, which quantifies the number of true positives and false positives. When prevalence increases, even though both the numerator and denominator are increasing from $[prev \times sensitivity]$, the denominator is also decreasing from $[(1 - prev) \times (1 - specificity)]$ decreasing (since as prevalence increases, the quantity $(1 - prev)$ decreases). Since sensitivity and specificity are constant, increasing prevalence has the effect of increasing the number of true positives (left term) and decreasing the number of false positives (right term). PPV increases when the number of true positives increases.

$$P(D|T^+) = \frac{\text{prevalence} \times \text{sensitivity}}{[\text{prev} \times \text{sensitivity}] + [(1\text{-prev}) \times (1\text{-specificity})]}$$

Table 2: Prevalence of Breast Cancer by Age Group

Age Group	Prevalence	PPV
30 - 40	0.0044	0.070
40 - 50	0.0147	0.202
50 - 60	0.0238	0.293
60 - 70	0.0356	0.386
70 - 80	0.0382	0.403

The following R chunk demonstrates how R can be used as a calculator; the code calculates PPV through applying Bayes' rule.

```
#calculations
prev = 0.0382
sens = 0.85
spec = 0.95

numerator = prev*sens
denominator = prev*sens + (1-spec)*(1-prev)

ppv = numerator/denominator
ppv
```

```
## [1] 0.4030536
```

Additional Comments:

The relationship between PPV and prevalence has very important implications in breast cancer and in screening populations for diseases. A paper published on February 9, 2017 (yes, just last week) in *Cancer Epidemiology Biomarkers & Prevention* suggests that women who experience a false positive mammogram are less likely to return for their regularly scheduled mammogram than women who receive a true negative. False positives are usually confirmed with a biopsy, and this unpleasant experience seems to have a negative influence on health behavior. You may be able to view the article through the Harvard Library online resources; it had not yet been posted on the morning of February 9.

More generally, small values of PPV in diseases with low prevalence make it impractical and sometimes unethical to screen for diseases with low prevalence. In the early days of the HIV epidemic in the US, there was considerable support for screening everyone for HIV. The epidemic was real, but the number of HIV+ individuals in the US was very small, and false positive tests might have led to the loss of employment or other stigma. Fortunately, mass screening was never implemented. It is possible to 'increase' prevalence by restricting the screened population to individuals at high risk for a disease, because the prevalence in the high risk population would be higher. Viewed this way, prevalence is a conditional probability of disease, given membership in a high risk population.

   d) Suppose that two new mammogram imaging technologies have been developed which can improve the PPV associated with mammograms; one improves sensitivity to 99% (but specificity remains at 95%), while the other improves specificity to 99% (while sensitivity remains at 85%). Which technology offers a higher increase in PPV? Explain your answer.

The technology that raises specificity to 99% offers a higher increase. For example, see the below code that tests how PPV changes for the 70 - 80 age group, as compared to the current PPV of 0.403; while raising sensitivity only increases PPV to 0.440, raising specificity increases PPV to 0.771. Since prevalence of the disease only ranges between less than 1% to at most ~ 4%, most of the people tested do not have breast cancer. This implies that the low PPV is largely due to the high number of false positives. This value is related to the specificity. If specificity increases, this increases the number of true negatives and decreases the number of false positives. Increasing sensitivity can potentially increase the number of true positives detected and raise PPV, but this does not have as strong an effect.

```
#test high sensitivity
prev = 0.0382
sens = 0.99
spec = 0.95
numerator = prev*sens
denominator = prev*sens + (1-spec)*(1-prev)
ppv = numerator/denominator
ppv
```

```
## [1] 0.4402151
```

```
#test high specificity
prev = 0.0382
sens = 0.85
spec = 0.99
numerator = prev*sens
denominator = prev*sens + (1-spec)*(1-prev)
ppv = numerator/denominator
ppv
```

[1] 0.7714788

5. Prostate-specific antigen (PSA) is a protein produced by the cells of the prostate gland. Blood PSA level is often elevated in men with prostate cancer, but a number of benign (not cancerous) conditions can also cause a man's PSA level to rise. The PSA test for prostate cancer is a laboratory test that measures PSA levels from a blood sample. The test measures the amount of PSA in ng/ml (nanograms per milliliter of blood).

The sensitivity and specificity of the PSA test depend on the cutoff value used to label a PSA level as abnormally high. In the last decade, 4.0 ng/ml has been considered the upper limit of normal, and values 4.1 and higher were used to classify a PSA test as positive. Using this value, the sensitivity of the PSA test is 20% and the specificity is 94%.

The likelihood that a man has undetected prostate cancer depends on his age. This likelihood is also called the prevalence of undetected cancer in the male population. The following table shows the prevalence of undetected prostate cancer by age group, where age is measured in years. Prevalence is measured as a proportion of the population. For instance, the third row shows that 6.0% of males age 61 - 70 have undetected prostate cancer.[2]

Table 3: Prevalence of Prostate Cancer by Age Group

Age Group	Prevalence	PPV	NPV
< 50	0.001	0.003	0.999
50 - 60	0.020	0.064	0.983
61 - 70	0.060	0.175	0.948
71 - 80	0.100	0.270	0.914

a) Calculate the missing PPV and NPV values, using any method.

R can be used as a calculator to efficiently calculate PPV and NPV according to the following equations:

$$\text{PPV} = P(D|T^+) = \frac{P(D) \cdot P(T^+|D)}{P(T)} = \frac{P(T^+|D) \cdot P(D)}{[P(T^+|D) \cdot P(D)] + [P(T^+|D^C) \cdot P(D^C)]}$$

$$\text{NPV} = P(D^C|T^-) = \frac{P(D^C) \cdot P(T^-|D^C)}{P(T^-)} = \frac{P(T^-|D^C) \cdot P(D^C)}{[P(T^-|D^C) \cdot P(D^C)] + [P(T^-|D) \cdot P(D)]}$$

---

[2]The numbers are only approximate, and vary by country.

```
prevalence = c(0.001, 0.020, 0.060, 0.100)
sensitivity = rep(0.20, 4)
specificity = rep(0.94, 4)

ppv.numerator = prevalence*sensitivity
ppv.denominator = ppv.numerator + (1 - prevalence)*(1 - specificity)
ppv = ppv.numerator/ppv.denominator
ppv
```

```
## [1] 0.003325574 0.063694268 0.175438596 0.270270270
```

```
npv.numerator = (1 - prevalence)*specificity
npv.denominator = npv.numerator + (prevalence)*(1 - sensitivity)
npv = npv.numerator/npv.denominator
npv
```

```
## [1] 0.9991488 0.9829279 0.9484757 0.9136069
```

b) Describe any observable trends in the PPV and NPV values.

As prevalence of prostate cancer increases by age group, PPV also increases. However, with rising prevalence, NPV decreases.

c) Explain the reason for the trends in part b), in language that someone who has not taken a statistics course would understand.

The probability that a man has prostate cancer, given a positive test, necessarily increases as the overall probability of having prostate cancer increases. If more men have the disease, the chance of a positive test result being a true positive increases (and the chances of the result being a false positive decreases). The decreasing NPV values follow similar logic: if more men have the disease, the chance of a negative test being a true negative decreases (and the chances of the result being a false negative increases).

d) The cutoff for a positive test is somewhat controversial. Explain how lowering the cutoff for a positive test from 4.1 ng/ml to 2.5 ng/ml would affect sensitivity and specificity. The answer to this part does not use any of the calculations in the previous parts of the problem.

Lowering the cutoff for a positive test would result in more men testing positive, since men with PSA values 2.5 ng/ml to 4.1 ng/ml were not previously classified as testing positive. Since the sensitivity of a test is the proportion who test positive among those who have disease, and the number with disease does not change, the proportion will increase, except in the rare and unlikely situation where the additional positive tests are among only men without the disease.

The effect on sensitivity can be seen by creating a table of a hypothetical population of 10,000 men. The table below corresponds to an overall prevalence of 0.05, sensitivity of 0.20, and specificty of 0.94. Note that since prevalence remains constant (in a given age group), the totals for D and D^C must remain the same.

	D	D^C	Total
T^+	100	570	670
T^-	400	8,930	9,330
Total	500	9,500	10,000

The following explanation is based on the idea that men with high PSA levels (i.e. between 4.1 ng/ml and 2.5 ng/ml) are far more likely to have prostate cancer. Thus, it is unlikely that an increase in positive tests will completely consist of false positives; it is also unlikely that all men who previously tested negative did not have prostate cancer.

– Suppose the number of positive tests (T^+, Total) were to increase to 700. Unless all of the additional 30 positive tests are false positives, the 100 in the upper left cell (T^+, D) must increase, and the number of false negatives (T^-, D) must decrease (since the total D must remain constant). This causes test sensitivity ($P(T^+|D)$), the ratio of the upper left cell to 500), to increase.

– The number of negative tests (T^-, Total) decreases, since a man's PSA level would now have to be below 2.5 ng/ml to be classified as testing negative. Unless all of the 30 previously negative tests were true negatives, the 8,930 in the middle cell (T^-, D^C) must decrease, and the number of false positives (T^+, D^C) must increase (since the total D^C must remain constant). This causes test specificity ($P(T^-|D^C)$), the ratio of the middle cell to 9,500), to decrease.

Alternative Justification. If the assumption is made that all additional positive tests are false positives, changing the 570 in the upper middle cell to 600, then the number of true negatives (T^-, D^C) decreases. This has no effect on test sensitivity and decreases test specificity.