# Module 4: Evaluation and Use of Diagnostic Tests

Contributors: Scott Wells, Amy Kinsley, Sandra Godden

In this module, you will explore how to understand how diagnostic and screening tests are evaluated, applied, and interpreted in individual animal (clinical medicine) settings and group-level (population medicine) settings, including the following:

1. How do I reach a diagnosis?
2. What are different types of tests?
3. How to evaluate the performance of diagnostic tests? Sensitivity and Specificity.
4. How to identify abnormal diagnostic test results?
5. Can I trust these test results? Positive and Negative Predictive Values.

## How do I reach a Diagnosis?

Clinical epidemiology involves the study of those things that help us arrive at an accurate and efficient diagnosis, prognosis, and treatment in the best interest of the patient. Diagnostic tests are one important component of clinical epidemiology, and considers: selection, evaluation, use, and interpretation of the test.

Three Maxims to remember:

* Common diseases occur commonly.
* Uncommon manifestations of common diseases are more common than common manifestations of uncommon diseases.
* No disease is rare to the patient who has it.

The Diagnostic Process is “the crucial process that labels patients and classifies their illnesses, that identifies (and sometimes seals!) their fates or prognoses, and that propels us toward specific treatments in confidence (often unfounded) that they will do better than harm” (Clinical Epidemiology, Sackett et al., 1991). Note: your ‘patient’ could be an individual animal or a herd.

### Strategies of Clinical Diagnosis (Sackett et al., 1991):

1. **Arborization or Algorithm**: Follow predefined decision tree. Extremely logical.
2. **Complete, exhaustive**: Complete diagnostic work-up which creates a comprehensive list of all problems and all possible differential causes, then proceeds to sift through the data for a final diagnosis. Disadvantage: is exhaustive, costly, time consuming, and inefficient. For example, there are 372 differential causes of recumbency in the bovine and 77 differential causes of pruritus in the feline.
3. **Hypothetico-deductive strategy**: “The formulation, from the earliest clues about the patient, of a short list of potential diagnoses or actions, followed by the performance of those clinical (history and physical exam) and paraclinical (laboratory, radiograph, etc.) maneuvers that will best reduce the length of this list.” (Sackett et al., 1991).

*Figure 1: The process of generating a working or definitive diagnosis while using the hypothetico-deductive strategy in a case workup (below).* Questions posed in the history taking and both the general inspection and routine exams are tests that help rule out certain diagnostic hypotheses, thus shortening our list of potential diagnoses. Response to treatment is also a very common test. Tests in the traditional sense (laboratory, radiographs, ultrasound, etc.) are often not required for diagnosis.

1. **Pattern Recognition**: Based on experience and previous cases in the clinician’s memory. This is reflexive, not reflective. The instantaneous realization that the patient’s presentation conforms to a previously learned picture (or pattern) of disease). Example: A dog with a papulo crustous pruritic dermatitis primarily in the lumbosacral region that may also involve the thighs, flanks, and ventral abdomen of a dog. There could be other causes, but experience tells us the most likely diagnosis is flea-bite allergy dermatitis.

*Caution: You will occasionally “get burned” when relying frequently on this approach to diagnosis.*

Medical and Veterinary students will most commonly use the hypothetico-deductive approach. With increased education and experience, clinicians become more likely to generate the correct hypothesis, to generate it earlier, and to collect more pertinent information relating to the working hypothesis (Sackett, 1991). With time and experience, clinicians move away from using the hypothetico-deductive approach on a conscious level, and begin using it on a subconscious level, moving increasingly towards pattern recognition. When presented with uncommon or unfamiliar problems, clinicians may fall back on the complete exhaustive method.

Figure 1: Diagnostic Process for Evaluation of a Presenting Complaint: Individual or Herd Level (Adapted from Ruminant Field Services Group, Ontario Veterinary College)



## What are the different types of tests?

* ***Diagnostic tests***: Often used in individual sick animals to establish or confirm a diagnosis, assess severity, determine prognosis, or assess progress or treatment response.
* ***Screening tests:*** Often used in apparently healthy populations or individual animals to detect carriers, subclinical disease, or seroprevalence (evidence of exposure). Often used for disease control, prevention, or eradication programs.
* ***Gold Standard tests*:** A test that is absolutely accurate in determining if disease is truly present or absent. For example, if we had a definitive test for Feline Immunodeficiency Virus, then the test would correctly identify all truly infected cats as positive and all truly uninfected cats as negative. With some diseases, gold standard tests to obtain a definitive diagnosis may include biological culture, biopsy, surgery, or postmortem. Although it would be ideal to use a gold standard test to make all diagnoses, many of these are elaborate, risky, expensive, or inappropriate in the timing of their application. For example, necropsy is too late as a diagnostic test for the patient (but not the population). In reality, there are very few true gold standard tests. Consequently, we often use an **accepted standard** test for defining an animal’s status for a target disorder. An accepted standard test may not be a “perfect” test, but is the best available. Other reasons for using other diagnostic or screening tests, even if they aren’t perfect, are improved speed, lower costs, or increased convenience, or reduced risk (less invasive).

## How to evaluate the performance of diagnostic tests? Sensitivity and Specificity

“Regardless of the reason for seeking diagnostic data, the overriding criterion to use when deciding which data to seek ought to be the *usefulness* of a given piece of diagnostic data to the clinician who seeks it and the patient who generates it.” (Sackett et al., 1991)

***Ask yourself “If I perform this test, will the result alter my diagnosis or course of action?” If the answer is ‘NO’, the test may not be worth performing.***

In selecting a test to make a medical decision, it is important to understand just how good or reliable these tests are, in comparison to the gold standard, or in comparison to other available accepted standard tests.

The usefulness of any given test (as compared to a gold standard or other accepted standard test) depends on:

* Apparent prevalence of disease
* True prevalence of disease
* Test Sensitivity (Se)
* Test Specificity (Sp)
* Predictive value of positive and negative test results (to be covered later)
* Other considerations, including:
* Cost
* Convenience
* Speed / Timeliness
* Safety (risk, invasiveness)

To assess how well a new diagnostic test performs, as compared to the gold standard test results, we must gather data on results of the diagnostic or screening test for comparison to the results of the gold standard (true state of nature). This assessment should be performed using a group of animals (sample population) similar to those animals with which you plan to use the test in real life (target population), with similar population characteristics (species, breed, age, sex, etc.) and similar in prevalence and spectrum of disease.

### **The 2-by-2 Table**

**Gold Standard Results**



n = Total number of individuals which were tested.

a + c = Total number of truly diseased individuals.

b + d = Total number of truly disease-free individuals.

a + b = Total number of individuals which tested positive.

c + d = Total number of individuals which tested negative.

### **Population Characteristics: True and Apparent Prevalence**

#### True Prevalence:

The proportion of individuals tested that are truly diseased according to the gold standard test

 = (a + c) / (a + b + c + d)

#### Apparent Prevalence:

The proportion of individuals tested that appear to be diseased based on the number of positive test results from our new test.

 = (a + b) / (a + b + c + d)

### **Sensitivity and Specificity: Test Characteristics of the New Test being Evaluated:**

#### **Sensitivity:**

True positive rate. The ability of the test to detect individuals with disease. Of the animals that truly are diseased, the proportion that test positive.
 = a / (a + c)

#### **Specificity:**

True negative rate. The ability of the test to correctly identify individuals who do not have the disease. Of the animals that truly don’t have the disease, the proportion that test negative.

 = d / (b + d)

#### **False Positive Rate:**

 Of the animals that truly don’t have the disease, the proportion that test positive.

 = b / (b + d)

#### **False Negative Rate:**

 Of the animals that truly do have the disease, the proportion that test negative.

 = c / (a + c)

#### **Test Accuracy:**

The proportion of all test results (both positive and negative) that are correct.

Refers to the validity of the test or the overall ability of the test to determine the true state of nature. Accuracy is affected by test sensitivity and specificity.

 = (a + d) / (a + b + c + d)

#### **Precision**:

 This refers to test consistency or **repeatability**. If you repeat the test on the same sample several times, how often would you get the same answer? A test that is not precise cannot be accurate. However, a test could possibly be precise and still not be accurate (i.e. the test could be consistently wrong).

**Note: Biochemical (or analytic**) **sensitivity** refers to the smallest concentration the test can detect (e.g. a hypothetical doping test can detect anabolic steroids in urine in concentrations as low as 10 ppb). Biochemical specificity refers to how narrow spectrum the test is in detecting only certain chemical agents (e.g., no cross-reactivity, a test might detect only beta-lactam antibiotic residues in milk will not cross react with macrolide antibiotic residues). Don’t confuse these terms with epidemiological sensitivity and specificity.

### **Case 4.1. Evaluation of a new Canine Heartworm test**Canine Heartworm disease is caused by a nematode infection, *Dirofilaria immitis*. Mosquitoes carrying infective heartworm larvae bite a dog to transmit the infection. The larvae grow, develop, and migrate in the body over a period of several months to become sexually mature male and female worms that reside in the heart, lungs, and associated blood vessels (e.g., caudal vena cava). The adult worms mate and release their offspring (microfilariae) into the bloodstream about 6-7 months after the initial infection. Heartworm infection can produce inflammatory processes, congestive heart failure, sudden collapse and, occasionally, death.

The following study (Courtney and Zeng, Vet. Parasit., 2001) evaluated one Heartworm Antigen Test Kit (SNAP®PF (IDEXX Laboratories) commonly used to diagnose Heartworm infection in dogs. The blinded study used the test kit to test serum drawn from 237 random source dogs. At necropsy (considered the gold standard test), 140 dogs were confirmed to have heartworm infection and 97 were confirmed heartworm free. 94 of the 140 truly infected dogs tested positive with the test. 95 of the 97 truly uninfected dogs tested negative with the test.

Fill in the two-by-two table and calculate the following population and test characteristics:

**Gold Standard Results**





# Answers

|  | Disease positive | Disease negative | Total |
| --- | --- | --- | --- |
| New test positive | 94 | 2 | 96 |
| New test negative | 46 | 95 | 141 |
| Total | 140 | 97 | 237 |

True prevalence in this study: = (a + c) / (a + b + c + d)

= 140 / 237 = 59.0%

Apparent prevalence in this study = (a + b) / (a + b + c + d)

= 96 / 237 = 40.5%

Sensitivity of test kit = a / (a + c)

= 94 / 140 = 67.1%

Specificity of test kit = d / (b + d)

= 95 / 97 = 97.9%

Overall accuracy of test kit = (a + d) / (a + b + c + d) = 189 / 237 = 79.4%

## How to identify abnormal diagnostic test results?

Identifying abnormal test results starts with understanding what is normal.

So far we have dealt with tests that have either a Yes or No outcome (e.g., disease positive or negative; pregnant or not pregnant). However, many tests, such as clinical biochemistry tests (e.g. serum calcium, phosphorus, liver enzyme concentrations) yield continuous numbers or a semi-quantitative outcome as results. Interpretation is based on lab-reported “Normal” Values. But what is “Normal”? How did the lab decide what is normal? There are different approaches that labs or researchers might use to decide what is ‘normal’:

### Approach 1. Setting Normal Limits based on Population Distribution:

* A sample of animals considered to be clinically (visibly) normal are evaluated with the test.
* The mean and standard deviations are calculated for the group results.
* Limits are set at 2 standard deviations (Mean +/- 2 standard deviations is considered the normal range).
* Thus, 95% of ‘apparently normal’ individuals will automatically always fall within the 2 standard deviations.



Normal Range = Mean +/- 2 Standard Deviations

Potential problems in using this approach to setting normal limits:

1. 5% of ‘apparently normal’ animals will always fall outside of this range (2.5% above and 2.5% below) and automatically be labeled as ‘abnormal’.
2. The apparent prevalence of disease would always be 5%.
3. ‘Normal’ or ‘typical’ in the population does not always mean ‘optimal’.
4. If we perform enough tests on an animal (e.g., HR, RR, CBC, chemistry profile, urinalysis, x-ray, blood gas analysis, MRI, etc.), sooner or later we will get an ‘abnormal’ result simply based on chance alone (e.g., 5% or 1 in 20 tests will fall into the ‘abnormal’ range), even if the patient is actually truly healthy. How do we differentiate a spurious test result from a truly sick animal?
* Re-test the animal and take the second test result as correct.
* Interpret the test result in light of all the other clinical data you have compiled, Does it make sense? Does it fit the clinical picture?
1. Was the sample group used by the lab for establishing normal limits representative of the current animal you are using the test on? (ie., species, breed, age, sex) Ask the lab.

### Approach 2. Setting Cut-Points by Considering Test Characteristics, Sensitivity, and Specificity

For many tests with continuous or semi-quantitative results, there is a normal distribution (bell-shaped curve) for normal and diseased animal populations, with overlap between the normal and diseased states. We are forced to decide on a cut-point to differentiate between diseased and healthy animals. To do this, we must make a trade-off between sensitivity and specificity.

For most tests, there is an inverse relationship between sensitivity and specificity. In Figure 2 (below) if we move the cutpoint to the left we will increase the sensitivity (detect more truly diseased animals and have fewer false negatives), but in so doing we will lower the specificity (will have more false positives). The opposite occurs if we move the cutpoint to the right (improve specificity but reduce sensitivity).



In the above scenario (Figure 2), when there is a trade-off between sensitivity and specificity, we can use a couple of different strategies to decide on the best cut-point:

#### Method 2a. Setting Cut-points using a Receiver Operator Characteristic (ROC) Curve

We can plot the sensitivity against the false positive rate (1-specificity) at different cut-point levels for a test to select the optimum cut-point for distinguishing between diseased versus not diseased animals.

##### **Case 4.2.** **Risk of abomasal displacement in cows**

For example, Dr. Thomas Geishauser determined the subsequent risk of **abomasal displacement (LDA) in Holstein dairy cattle** from serum AST levels obtained in the second week post-calving. He determined the sensitivity and specificity values for predicting LDA at various serum AST cut-off values (Am J Vet Res. 58:1216, 1997):



After totaling sensitivity and specificity we can see that their sum is maximized at a cut-off of ≥100 U/L of serum AST in the second week post-calving. This can be depicted in a Receiver Operator Characteristic Curve (ROC curve; see below).



With the ROC curve, the cut-off value of > 100 U/L is closest to the upper left hand corner of the plot, and so is the optimal cut-point. In our example, the cut-point of > 100 U/L resulted in the fewest gross number of diagnostic errors. When comparing two tests, the better test is the one with the greatest area under the curve (i.e., the curve most closely approaches the upper left-hand corner). Note that using the ROC method of selecting an optimal cut-point assumes that a false-positive is as just as bad as a false-negative (the two errors are weighted equally). However, this may not always be true for some diseases. In the latter case, we might use a different approach (see method 2b. below).

#### **Method 2b.: Setting Cut-Points based on Importance of Optimizing Either Test Sensitivity or Specificity**

In selecting the best cut-point, we need to consider:

1. What is the distribution of results among diseased and disease-free animals?
2. What is the prevalence of disease in the target population?
3. Why are we using the test? Is it to identify positive animals or disease free animals?
4. What are the potential costs or consequences of a false-negative or false-positive result? Which would we prefer to avoid?

From this, we can decide if we would prefer to maximize either Sensitivity (minimize false-negatives) or maximize Specificity (minimize false-positives), and then adjust the cut-point either up or down accordingly.

As an example, from Dr. Geishauser’s work, what cutpoint should we select for trying to predict cows that will develop a displaced abomasums?



How do we decide whether we need a sensitive or a specific test?

* Highly sensitive tests will produce few false-negatives.
* Highly specific tests will produce few false-positives.

What do we want to avoid more, a false-positive or a false-negative?

* What is the cost or risk of a false-negative?
* What is the cost or risk of a false-positive?

To make this decision, you must consider consequences to the animal (life or death, quality of life), consequences to the owner (emotional, financial / livelihood), and consequences to the herd (e.g., TB eradication program), and possibly even consequences to the industry (e.g., what is the consequence of missing a diagnosis for a highly contagious foreign disease such as foot and mouth disease?).

##### **SNOUT: When will we want to maximize Sensitivity?**

1. Sensitive tests rarely miss individuals with disease (few false-negative results)
2. **SnNout**: If a highly ***Sensitive*** test yields a ***Negative*** result, this is useful to **Rule-Out** the disease in that animal.
3. Most helpful when results are negative.
4. Highly sensitive tests are most useful when:
* Used in the early stages of a diagnostic workup to RULE-OUT other diseases.
* When we want to avoid false-negatives.
* When missing an infected animal (i.e., a false-negative result) could result in serious negative consequences to the animal, client or population:
* Example 1. Missed diagnosis for Foot and Mouth disease results in rapid spread to other farms.
* Example 2. Missed detection of ‘pregnancy’ in a cow that is already 180 days in milk could result in needless culling of the animal from the herd.
* Useful when the probability of disease is low (low prevalence) and the purpose is to discover disease (e.g., in an eradication program, use as a herd screening test to detect animals infected with bovine tuberculosis).

##### **SPIN: When will we want to maximize Specificity?**

1. Specific tests are rarely positive in the absence of disease (few false-positives)
2. **SpPin**: If a highly ***Specific*** test yields a ***Positive*** result, this is useful to ***Rule-In*** the disease in that animal.
3. Most helpful when results are positive.
4. ***Highly specific tests are most useful when:***
* Useful to confirm or Rule-In a diagnosis.
* Most helpful when results are positive.
* Critical when want to avoid false positive results that could lead to unnecessary harm to the individual or the herd, or cause significant emotional or financial costs to the client:

- Example 1: Unnecessary euthanasia of a pet

- Example 2: Required slaughter of an entire herd of elk if an animal among the herd tests positive for Tuberculosis.

## Can I trust these test results? – Positive and Negative Predictive Values

Sensitivity and specificity are important descriptors of how well a test performs. Any new diagnostic test should be independently and rigorously tested to describe the test characteristics (e.g., by independent (university) researchers). Clinicians can then use that information to make educated decisions about whether or not to adopt a test into their practice setting. However, after adopting a new test into our practice, as clinicians we may not have the luxury of having a “gold standard” easily available to check the accuracy of our new test (if it was, we wouldn’t be considering using this other test, would we?). As clinicians we wish to know, based on the test result attained, the probability that the animal has or does not have the disease in question. Can we trust the test result?

### **Case 4.3. Calculating and interpreting predictive values**

You work in a Florida practice, and get either a positive or negative test result on a dog when using the SNAP PF Heartworm antigen test kit. Knowing that false positive or false negative results are possible, can you trust that test result to be correct? In other words, what is the test’s predictive value? ***What is the probability that a test result reflects the true disease status?***

Use the data from the Canine Heartworm disease example to calculate the predictive values of both a positive or negative test result:

Info you already know from the previous Florida study (Courtney. 2001. J. Vet. Parasit):

* Total test population from study done in Florida = 237 animals
* 140 animals were found to be truly infected at necropsy
* 97 animals were found to be truly uninfected at necropsy
* Assume the prevalence in your client population is similar to the original study
* Test sensitivity = 67% (false-negative rate = 33%)
* Test specificity = 98% (false-positive rate = 2%)

|  | Disease Positive | Disease Negative |  |
| --- | --- | --- | --- |
| Test Positive | a 94 | b 2 | a + b 96 |
| Test Negative | c 46 | d 95 | c + d 141 |
|  | a + c 140 | b + d 97 | n 237 |

1. ***Predictive Value of a Positive Test (PV+ve)***

The proportion of test positive animals that are truly diseased.
or, The probability, given a positive test result, that the animal actually has the disease.

= a / (a + b)
=

1. ***Predictive Value of a Negative Test (PV-ve)***

The proportion of test negative animals that are truly not diseased.
or, The probability, given a negative test result, that the animal does not have the disease.

 = d / (c + d)

 =

1. Practice interpreting your results in full sentences that a client would understand.

**Answers**

1. ***Predictive Value of a Positive Test (PV+ve)***

The proportion of test positive animals that are truly diseased.
or, The probability, given a positive test result, that the animal actually has the disease.

= a / (a + b)

 = 94/96 = 97.9%

1. ***Predictive Value of a Negative Test (PV-ve)***

The proportion of test negative animals that are truly not diseased.
or, The probability, given a negative test result, that the animal does not have the disease.

 = d / (c + d)

 = 95/141 = 67.4%

1. Practice interpreting your results in full sentences that a client would understand.

Based on this data relating to Florida:

* + 98% of test-positive dogs are truly infected.
	+ 67% of test-negative dogs are truly uninfected.