# **Module 7: Evaluating Epidemiologic Associations and Their Relationship to Causation**

Contributors: Scott Wells, Amy Kinsley, Julio Alvarez

**Key Concepts**

1. Considering if epidemiologic associations are causal.
2. Evaluate epidemiologic associations between exposures and outcomes.

## **1. Differentiating Association from Causation:**

Koch’s Postulates (1884) provided the classic framework for identifying causes of infectious disease using criteria to be met before an agent (e.g., virus or bacteria) could be considered as the cause of a disease:

* The agent has to be present in every case.
* The agent has to be isolated from the affected individual and grown in pure culture.
* The agent has to cause disease when inoculated into a susceptible animal and the agent must then be able to be recovered from that animal and identified.

Unfortunately, problems arose when using Koch’s postulates to determine causation of disease. Some of these problems included:

* The requirement that a disease has only one cause
* The requirement that a specific cause should result in only one disease
* Diseases with multiple etiologic factors
* Multiple effects (i.e., diseases) from single causes
* Carrier states
* Non-agent factors (age, breed) that predispose to disease
* Quantitative causal factors

Many epidemiological studies carried out worldwide in veterinary medicine try to identify the **cause(s)** that lead to certain (typically undesired) health outcomes. We define **cause** as “***any factor that produces a change in the severity or frequency of the outcome***” (Dohoo, 2007).

The best way to demonstrate the causal relationship between an exposure (our cause ‘candidate’) and the outcome of interest (the disease) is to observe what would have happened to the exposed individuals if that exposure did not take place (the *counterfactual group*). Of course, this is not possible, and for this reason, we have to instead settle for a set of different individuals that will compose the **control group**(the unexposed group that ideally will be as comparable to the exposed group as possible other than the exposure of interest).

In an **experimental study,** the investigators usually have control over the individual animals that enter the study, and therefore the study animals can be **randomized** (i.e., randomly allocated) into exposure and control groups. Through this randomization, we expect that these two groups (those exposed to the treatment and those not exposed to the treatment) will end up being as equivalent as possible, other than the treatment exposure (particularly if the sample size is large enough). In addition, since the investigators usually administer the exposure, we expect the outcome will take place after the individuals were exposed (since any individual that had experienced the outcome before the experimental study started would not be included in the study).

For these reasons, we typically rank experimental studies above observational studies in the hierarchical pyramid of quality of scientific evidence (this will be shown in the next course module), although of course experimental studies are not necessarily perfect (and there are certain things that cannot be studied through experimental studies).

Due to the multiple conditions in which a study (experimental or observational) can be performed, it is common that results across studies differ or are inconclusive, and, for this reason, proving that a certain exposure is causal (leads to an outcome) can be very difficult.

In this context, the application of a set of criteria can help to define the evidence pointing towards or against a causal relationship between a potentially causal exposure and a health outcome. Among these, the **Bradford Hill Principles for Causation** have been widely used, particularly in a public health context. These principles include (see **Stevenson’s text, page 56** for more details):

### Bradford Hill Principles for Causation

1. Strength (of association) – Often measured using relative risks and odds ratios.
2. Consistency (reproducibility)
3. Specificity
4. Temporality (cause precedes effect)
5. Dose-response relationship (greater exposure leads to greater effect)
6. Plausibility and Coherence (from epidemiology to laboratory)
7. Experimental evidence
8. Analogy

Note that these set of criteria can be considered a useful aid when interpreting results from one or multiple studies, but often they may not be all fulfilled, and can be interpreted differently by different individuals, so none of them should be considered indisputable evidence for or against the hypothesis of a causal relationship between exposure and effect.

Proving causation is particularly complex due to the usual involvement of more than one cause (or exposure) in the process that leads to the outcome (for example, a disease). Of course, not all causes will be equally important, and, depending on the criteria, causes can be classified as:

#### Necessary cause:

An exposure needed for the disease to occur. If the exposure does not happen, there is no outcome, so that the exposure will have always taken place when the outcome occurred.

#### Sufficient cause:

An exposure that leads to disease. Once the exposure has taken place, the outcome always happens. More often than not, there is not a single sufficient cause but a combination of exposures that altogether become a sufficient cause, and each of these are termed a component cause.

#### Component causes:

For example, for a given animal, *being in poor body condition* and *housed in a poorly ventilated barn* and *with other infected animals* may be an example of a sufficient cause for certain respiratory diseases. In this case, the sufficient cause is formed by combination of the three component causes.

Similarly, causes may also be classified into:

#### Direct cause

(when there is no additional factor intervening in the development of the outcome) or

#### Indirect cause

(when effects on the occurrence of the outcome are mediated through other factors).

The following exercise can help to illustrate how to evaluate evidence for causation.

#### **Case 7.1. Does Bovine leukemia virus (BLV) cause breast cancer in women?**

Human breast cancer is the most frequent cancer in women globally, with 1.7 million new cases diagnosed annually. This is the 4th highest cause of death in women in developed countries, and 6th highest in developing countries. Breast cancer is considered to be a multifactorial disease with several known risk factors: estrogen, radiation, age, race, number of births, obesity after menopause, and certain genetic mutations. While human breast cancer is not usually considered to have an infectious etiology, many viruses are associated with human cancers, and approximately 12% of cancers (at this time) have a viral etiology.

In comparison, bovine leukosis is a cancerous disease of dairy and beef cattle, with tumors often found in the heart, uterus, abomasum, and external lymph nodes of cattle. This disease in cattle is caused by **Bovine leukemia virus (BLV)**, a retrovirus that targets lymphocytes. BLV infection occurs at a high prevalence in dairy cattle in the US and other countries, with an estimated seroprevalence of approximately 45% shown in several studies.

Based on the high prevalence in dairy cattle and potential exposure to people from consumption of dairy and beef products, some people have hypothesized that BLV could be a cause of breast cancer in women. This question was evaluated using the Hill criteria for causation in a recent publication titled: Does BLV meet the criteria for causation of human breast cancer? (Reference: Cuesta et al. 2018. Journal of Mammary Gland Biology and Neoplasia 23:101–107). <https://pubmed.ncbi.nlm.nih.gov/29777406/>

##### Using the Hill criteria for causation:

1. **Strength (of association):** The higher the magnitude of the association, typically measured using a relative risk (RR) or odds ratio (OR), the more likely the association is to be causal.

‘A similar analysis, performed in 2015, analyzed viral gene tax expression using in situ PCR. The target gene was amplified in 29% of the controls, 38% of the premalignant lesions and 59% of malignant samples. The **odds ratio** of the association between the presence of BLV DNA and breast cancer was 3.07, a value similar to that obtained for other risk factors such as hormonal exposure, number of births and lifestyle.’

1. **Consistency (reproducibility):** Indicated by more than one (or multiple) different studies identifying an association between the risk factor and the outcome.

‘Recently, a case control study of women in Texas supports the results found in previous studies showing an association of BLV DNA with breast cancer.’

1. **Specificity:** An association is more likely to be causal if a single exposure generally causes a single disease.

While the criterion for specificity is that one exposure causes one disease, more broadly, if there is an association in a specific population group with the same exposure, this provides evidence for the argument for causation.  The evidence is not clear in this case.

1. **Temporality:** An association is more likely to be causal if the exposure comes before the effect in time.

‘Recently, a study conducted in Australia detected the presence of a BLV gene before and after breast cancer development. These results showed that viral DNA could be detected in this tissue many years before the onset of the disease.’

1. **Dose-response relationship:** An association is more likely to be causal if the greater the exposure, the greater the effect (higher RR or OR).

As stated above, ‘Recently, a case control study of women in Texas supports the results found in previous studies showing an association of BLV DNA with breast cancer. They detected a fragment of the tax gene in 20% of benign, 34% of pre-malignant, and 57% of malignant tissues.’

‘There is a correlation between bovine milk and meat consumption and breast cancer. Countries with high consumption of these products have a higher prevalence of breast cancer than countries that consume fewer products derived from cattle. Moreover, women with lactose intolerance have a lower prevalence of breast cancer than other women and other family members.’

1. **Plausibility and coherence:** An association is more likely to be causal if the causal interpretation fits with known facts of the natural history and biology of the disease (from epidemiology to laboratory).

‘The exact way by which the virus might be transmitted to humans is still unknown. It was proposed that unpasteurized milk or undercooked meat could be a way in which the virus enters the body. It has been reported that the genetic material corresponding to the viral gene gag is present in fresh milk and raw meat.  However, it is well known that pasteurization and cooking inactivates this virus.’ ‘The Retroviridare family includes HIV, HTLV-1 and BLV, which share structural and functional homology. Both HIV and HTLV had been classified as a group I carcinogenic virus by the International Agency for Research on Cancer. BLV belongs to the same family as these human carcinogenic viruses. Thus, it is possible that BLV could be carcinogenic in humans. This possibility has not yet been studied thoroughly.’

1. **Experimental evidence:** An association is more likely to be causal if supported by experimental data (or data from clinical trials).

‘For an infection to occur, there must be a receptor that allows the virus to enter the cell. In the case of BLV, scientists have not yet identified the receptor. It is believed that the receptor is the δ subunit of bovine adaptor protein complex 3 (AP-3). Humans have four types of AP complexes that are widely distributed in the body and have an important role in intracellular transport. The existence of a human homolog of bovine BLV receptor means that there is a chance that BLV can infect this species. In fact, scientists were able to infect human cell lines with the virus in vitro.’

1. **Analogy:** An association is more likely to be causal if a similar relationship has been observed with another exposure and/or disease.

‘Lastly, following analogy criteria, BLV and HTLV structural and functional similarity can let scientists consider the possibility that BLV could affect humans in some way.  BLV and HTLV are both members of the Deltaretrovirus genus, characterized by the presence of a unique sequence called pX region located between the env gene and the 3’LTR. This sequence encodes regulatory proteins that continuous contact between BLV and humans could lead to a virus adaptation to this new host.’

As you can observe, evaluating causation from observational studies can be quite challenging. It can sometimes be difficult to determine if the criteria for causation are met.

How many of Hill's criteria for causation do you believe were met?

Do you believe BLV causes breast cancer in women?

## **2. Evaluate epidemiologic associations between exposures and outcomes**

### Quantifying the strength of an association

In veterinary practice, we often need to make decisions about the association between an exposure and a health outcome. Example situations include:

* Effectiveness of a treatment (exposure) to improve a chronic health condition (outcome).
* Effectiveness of a vaccination program (exposure) to prevent new cases of infectious disease (outcome).
* Importance of outdoor housing (exposure) as a risk factor for vector borne disease (outcome).

As discussed in the causation section above, one of the key criteria for evaluating causation is the **strength of an association** between an exposure (e.g., a treatment) and an outcome (e.g., a disease). We typically consider those exposures that are strongly associated with an outcome as more relevant.

To measure the strength of the association, we usually compare the frequency of the outcome in the exposed group to the frequency of the outcome in the non-exposed group. If the frequency of the outcome is much larger in the exposed group (relative to the non-exposed group), we have an indication that the exposure was related (associated) with at least a proportion of the outcomes that occurred.

Assuming that the outcome of interest is a disease, the measure of association we will use will depend on the way we have measured disease occurrence (i.e., incidence or prevalence).

We can create a 2x2 table to evaluate the strength of epidemiologic association as follows:

|  | **Diseased** | **Non-diseased** | **Total** |
| --- | --- | --- | --- |
| **Exposed** | a | b | a+b |
| **Non-exposed** | c | d | c+d |
| **Total** | a+c | b+d | **N** |

In terms of exposures, we typically place the exposed group in the top row of the table and the non-exposed group in the bottom row.

In terms of the outcome, we place the animals with the health outcome of interest (diseased for example) in the left column and the animals without that outcome (non-diseased) in the right column.

**Note:** The organization of this table in this exact way is important. You could switch the columns, but in that case, you would need to change the way you interpret the results from the association, so it is much easier to understand these associations by putting data into the table in exactly this way.

**Another Note: This 2x2 table is different from the 2x2 table we used to evaluate diagnostic tests.** While it may appear similar, when evaluating associations (in this situation), the column headings are categories of an outcome (often disease or test-positivity) and the row headings are categories of exposures (exposed or non-exposed). Recall that in the 2x2 table to interpret diagnostic test performance, the column headings were infection status (or results from a gold standard test) while the row headings were the test to be evaluated.

There are two primary methods of comparing disease occurrence among exposed and non-exposed individuals, and these involve **Ratios** and **Differences**.

* **Ratios** can adopt values between 0 and infinity, and will express how many times the risk is higher (or lower) in the exposed group compared to the non-exposed group. This ratio = Risk in exposed group / Risk in non-exposed group.
	1. A value of 1 indicates no difference in the risk between groups
	2. A value greater than 1 indicates a higher risk in the exposed group compared to the unexposed group
	3. A value less than 1 indicates a lower risk in the exposed group compared to the unexposed group.
* **Differences** can adopt negative and positive values, and express the difference between the risks in the exposed and non-exposed groups. This difference = Risk in exposed group minus Risk in non-exposed group.
	1. A value of 0 means there is no difference in the risk between the groups.
	2. A value greater than 1 indicates a higher risk in the exposed group compared to the non-exposed group.
	3. A value less than 1 indicates a lower risk in the exposed group compared to the non-exposed group.

The most commonly used **measures of strength of association** are:

#### **Incidence risk ratio, often called Relative risk (RR)**

The RR is the ratio of the incidence in the exposed group divided by the incidence in the non-exposed group. Think of this as a ratio of incidence risks. Using the 2x2 table set up as shown above (with the exposed group in the top row), RR can be calculated as

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

As true for all ratios, RR can adopt values between 0 and infinity. A RR = 1 indicates no difference in risk between the exposed and non-exposed groups. A RR >1 indicates a higher level of risk in exposed group compared to the non-exposed group, and a RR<1 indicates a lower level of risk in exposed group compared to the non-exposed group.

Sometimes relative risk can also refer to the **Incidence Rate Risk Ratio**, that compares the incidence rates in the exposed and the non-exposed groups (and is interpreted in the same way as the incidence risk ratio). If you are confused about the difference between incidence risks (a proportion) and incidence rates (a ratio), this would be a good time to review the notes from Module 3 of this course to help your understanding.

#### **Prevalence ratio (PR)**

The PR is the ratio of the prevalence in the exposed group divided by the prevalence in the non-exposed group. The PR is calculated in the same way as the RR.

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

so the only difference between the RR and PR is the nature of the measures they are comparing (incidence data or prevalence data).

#### **Odds ratio (OR)**

The OR is the ratio of the odds of disease in the exposed group divided by the odds of disease in the non-exposed group. The OR can be calculated as

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

which can be simplified to

OR = (a\*d) / (b\*c)

The OR is interpreted similarly to the RR, in that a value less than 1 indicates lower risk in the exposed group, a value greater than 1 indicates higher risk in the exposed group, and a value equal to 1 indicates similar risk in both groups. In this case, however, the OR is an indication of the odds of disease in the exposed group compared to the odds of disease in the non-exposed group.

#### **Risk difference (RD) or Attributable risk**

The RD is calculated as the difference between the risk in the exposed group and the risk in the non-exposed group. It expresses the increase (or decrease, if below 0) risk in the exposed group attributable to the exposure.

RD = Risk of disease in exposed group - Risk of disease in non-exposed group

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

RD = Risk of disease in exposed group - Risk of disease in non-exposed group

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

Next, let’s use an example to demonstrate how to use these values to evaluate the strength of associations.

### **Case 7.2. Intensively managed and backyard pig farms and the incidence of Porcine Reproductive and Respiratory Syndrome (PRRS)**

We have followed two groups of 100 intensively managed and 100 backyard pig farms for one month to detect new cases of PRRS:

* 100 intensively managed farms: 6 farms became PRRS positive
* 100 backyard farms: 18 farms became PRRS positive
1. What was the incidence risk of PRRS in intensively managed farms?
2. What was the incidence risk of PRRS in backyard farms?
3. Create a 2x2 table to evaluate the association between PRRS positivity at the farm level and management system (intensively managed or backyard).

|  | PRRS positive | PRRS negative | **Total** |
| --- | --- | --- | --- |
| Backyard farms |  |  |  |
| Intensively managed |  |  |  |
| **Total** |  |  |  |

1. What was the relative risk of PRRS positivity among these farms?  Which group had higher risk?
2. What was the attributable risk of PRRS positivity due to intensive vs backyard management?
3. What was the odds ratio for the association between management system and PRRS positivity in these data?

**Answers:**

1. What was the incidence risk of PRRS in intensively managed farms?

6/100 = 6%

1. What was the incidence risk of PRRS in backyard farms?

18/100 = 18%

1. Create a 2x2 table to evaluate the association between PRRS positivity at the farm level and management system (intensively managed or backyard).

|  | PRRS positive | PRRS negative | **Total** |
| --- | --- | --- | --- |
| Backyard farms | 18 | 82 | 100 |
| Intensively managed | 6 | 94 | 100 |
| **Total** | 24 | 176 | **200** |

1. What was the relative risk of PRRS positivity among these farms?  Which group had higher risk?

RR = 0.18 / 0.06 = 3. There was 3 times greater incidence of PRRS positivity among backyard farms (exposed) compared to intensively managed farms (unexposed).

1. What was the attributable risk of PRRS positivity due to intensive vs backyard management?

Attributable risk = 0.18 - 0.06 = 0.12. There was 12% higher risk of PRRS positivity among backyard farms (exposed) compared to intensively managed farms (unexposed).

1. What was the odds ratio for the association between management system and PRRS positivity in these data?

OR = (18 \* 94) / (6 \* 82) = 3.43. There was 3.43 times greater odds of PRRS positivity among backyard farms (exposed) compared to intensively managed farms (unexposed).

## **3. Evaluate statistical associations between exposures and outcomes and provide inference for these associations**

Even if no true causal association exists between an exposure and a health outcome, the risks of disease among the exposed and non-exposed groups will never typically be identical and will instead be somewhat different due to chance. In order to establish if the strength of an association (RR, PR, OR, RD) is truly different from 1.0 (RR, PR, OR) or 0 (RD), we use statistical tests that will compare the observed value with what would be expected if no differences truly existed. From classic statistics, this corresponds to a **null hypothesis** of no difference between risks of disease in the exposed vs. n on-exposed groups beyond that due to chance alone.

### **P-values**

Most classical statistical tests will generate a **p-value** that expresses the **probability of observing the data generated if the null hypothesis was true** (in this context, no difference in risk of disease between groups).

Conventionally, p=0.05 is considered a universally accepted threshold, so that p<0.05 (there is less than a 5% chance that the null hypothesis is true) provides evidence that we feel safe rejecting the null hypothesis, and instead consider the difference in risk of disease observed between groups is truly significant (exposure is associated with the disease).

**Note: This universally accepted threshold (p<0.05) is only a convention and not an absolute criteria for decision-making.** In addition to establishing the statistical significance of the association between exposure and disease), it is also important to evaluate the **biological significance of the association**. A RR of 1.05 may be statistically significant (statistically different from 1.0), especially with very large sample sizes, but may not be biologically significant if the increase in the risk is too low to be relevant biologically.

### **Chi-square test of independence (or association)**

For 2x2 tables (the primary format for epidemiologic evaluation used in this course), a simple method of evaluating the statistical significance of the association is the **Chi-square test of independence (or association)**. For this statistical test, assumptions are:

* The data in the cells are frequencies (counts of animals) rather than percentages or other data transformations.
* The categories of the variables are mutually exclusive (e.g., animals are either exposed or non-exposed).

The Chi-square test of association can be calculated manually if desired by comparing the observed results to the expected results, given no association between the exposure and disease status, as follows.

**Observed information** (the data you collected):

|  | Diseased | Non-diseased | **Total** |
| --- | --- | --- | --- |
| Exposed | a | b | a+b |
| Non-exposed | c | d | c+d |
| **Total** | a+c | b+d | **N** |

**Expected information** (assuming no association between exposure and disease):

|  | Diseased | Non-diseased | **Total** |
| --- | --- | --- | --- |
| Exposed | a | b | a+b |
| Non-exposed | c | d | c+d |
| **Total** | a+c | b+d | **N** |

The **chi-square test-statistic** is the sum of the squared differences between the observed and expected values. This value is compared to the distribution in a Chi-square table (Table below) to estimate the likelihood the observed values would have occurred in the population, assuming no association between the exposure and disease outcome (the p-value).

### **Case 7.3. Intensively managed and backyard pig farms and the incidence of Porcine Reproductive and Respiratory Syndrome (PRRS)**

As an example, we will again use the case evaluating the association between PRRS positivity at the farm level and management system (intensively managed or backyard).

The **observed data** you already evaluated:

|  | PRRS positive | PRRS negative | **Total** |
| --- | --- | --- | --- |
| Backyard farms | 18 | 82 | 100 |
| Intensively managed | 6 | 94 | 100 |
| **Total** | 24 | 176 | **200** |

Next, you need to generate the **expected data**, given no association between farm-level management system and PRRS positivity. To do this, you keep the row and column totals in the 2x2 table and change the internal values in the following way.

|  | PRRS positive | PRRS negative | **Total** |
| --- | --- | --- | --- |
| Backyard farms | a | b | 100 |
| Intensively managed | c | d | 100 |
| **Total** | 24 | 176 | **200** |

First, calculate the expected value in the ‘a’ cell by multiplying the row total by the column total and dividing by the grant total (100\*24/200 = 12). Repeat for the cell’s ‘b’, ‘c’, and ‘d’. This will result in the following expected table, given no association between farm-level management system and PRRS positivity.

|  | PRRS positive | PRRS negative | **Total** |
| --- | --- | --- | --- |
| Backyard farms | 12 | 88 | 100 |
| Intensively managed | 12 | 88 | 100 |
| **Total** | 24 | 176 | **200** |

The next step is to calculate the squared differences between the observed (O) and expected (E) values in each cell, then sum these squared differences.



Finally, compare this value to a Chi-square distribution with one degree of freedom (used in this case because from a 2x2 table).

Chi-square statistic = (18-12) ^2/12 + (82-88) ^2/88 + (6-12) ^2/12 + (94-88)^2/88

 = 6.8 with 1 degree of freedom (see Table below)

The p-value (probability of exceeding this value given no association between the exposure and disease outcome) is about 0.01. This very low p-value provides evidence that suggests there is an association between farm management and PRRS positivity in these data (there is only about 1% probability that the Null Hypothesis of no association between the outcome and disease is true).



### **Case 7.4. Incidence of leishmania in dogs living in a small town associated with walking dogs close to a park.**

You are concerned about the incidence of leishmania in dogs living in a small town. Leishmaniasis is transmitted by small biting sand flies in tropical areas of the world (fortunately not in Minnesota, at least yet). Since a preferred habitat of this parasite is near water, the risk factor of interest for evaluation was walking dogs close to a park with a lake. In a population of 100 dogs, the incidence risk over a 1 month period was 12%.  Of the dogs that developed the disease, 67% were walked close to the park (the risk area of concern). Of the non-diseased dogs, only 12 reported walking near or in the park.

|  |  |  | **Total** |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
| **Total** |  |  |  |

1. What was the incidence risk of clinical leishmaniasis in dogs walked close to the park?
2. What was the incidence risk of clinical leishmaniasis in dogs ***not*** walked close to the park?
3. What is the incidence risk ratio (Relative risk) for the association between clinical leishmaniasis and walking the dog close to a park?
4. What is the interpretation of this RR value in real world terms?
5. Is this association statistically significant?
6. What was the risk difference?

**Answer:**

|  | Clinical signs | No signs | **Total** |
| --- | --- | --- | --- |
| Walked near park | 8 | 12 | 20 |
| Not walked near park | 4 | 76 | 80 |
| **Total** | 12 | 88 | **100** |

1. What was the incidence risk of clinical leishmaniasis in dogs walked close to the park?

Incidence risk = 8 / 20 = 40%

1. What was the incidence risk of clinical leishmaniasis in dogs *not* walked close to the park?

Incidence risk = 4 / 80 = 5%

1. What is the incidence risk ratio (Relative Risk) for the association between clinical leishmaniasis and walking the dog close to a park?

Relative risk of clinical signs given walked near part = 0.40 / 0/05 = 8

1. What is the interpretation of this RR value in real world terms?

From these data, the risk of incidence of clinical signs was 8 times higher in dogs that walked near the park compared to dogs not walking near the park.

1. Is this association statistically significant?

From the Chi-square test of independence spreadsheet, the Chi-square test statistic = 18.56, which is statistically significant at the p=0.05 level of significance.

1. What was the risk difference?

The risk difference between dogs walking near the park (exposed) and dogs not walking near the park (nonexposed) = 40% - 5% = 35%