



---

**The Young Epidemiology Scholars Program (YES) is supported by  
The Robert Wood Johnson Foundation and administered by the College Board.**

## **Measures in Epidemiology**

**Manuel Bayona**

Department of Epidemiology  
University of North Texas  
Fort Worth, Texas

and

**Chris Olsen**

Mathematics Department  
George Washington High School  
Cedar Rapids, Iowa

# Contents

Introduction . . . . .	3
Teacher’s Notes on Measures in Epidemiology . . . . .	5
Lesson Plan . . . . .	7
Part I: Measures of Frequency . . . . .	9
Worksheet 1: Practice Problems in Prevalence and Incidence (Student Version) . . . . .	13
Worksheet 1: Practice Problems in Prevalence and Incidence (Teacher’s Answer Key) . . . . .	18
Part II: Measures of Association . . . . .	23
Worksheet 2: Practice Problems in Measures of Association (Student Version) . . . . .	30
Worksheet 2: Practice Problems in Measures of Association (Teacher’s Answer Key) . . . . .	32
Quiz: Measures in Epidemiology (Student Version) . . . . .	35
Quiz: Measures in Epidemiology (Teacher’s Answer Key) . . . . .	38

# Introduction

Epidemiology is the study of health-related events in populations and the application of this study to the understanding and control of health problems. A common example of a health-related event is disease. When individuals in a population are coming down with, say, measles, epidemiologists are concerned with establishing what individuals and groups have the disease, who is at risk of getting the disease, and why they might be at risk. It is hoped that this knowledge will enable health officials to understand how to control outbreaks of the disease.

Although the word epidemiology sounds an awful lot like epidemic, epidemiology is not solely concerned with epidemics. All health-related events are of interest to the epidemiologist, and the principles of epidemiology are applied across the board. When the term health-related event is used, it may refer to a contagious disease, an environmental situation that is dangerous to health, or even accidental injuries.

In searching for explanations about health-related events, epidemiologists are typically concerned with the following questions:

1. Which individuals have experienced the event?
2. When did they experience the event?
3. Where are the individuals who have experienced the event?
4. What environmental factors are associated with the event?

Two characteristics of scientific disciplines such as epidemiology are:

1. A specific language for purposes of communication about its subject
2. Quantitative measures for describing and analyzing the targets of study

Epidemiologists will use not only existing common words and phrases in very specific ways but also uncommon words and phrases. If necessary they may even invent some new terminology now and then. With this terminology epidemiologists make distinctions about health events that nonepidemiologists do not usually have to make. Quantitative measures are necessary to describe important aspects of health-related events. In this module we will learn about two fundamental concepts of epidemiology and the measurements used to describe these concepts numerically.

First, we will be concerned with measuring the frequency of health-related events. The frequency of an event is a basic building block when comparing the health status of different populations, different subpopulations within a population and populations at different times. Do French citizens have less risk for heart disease? Are males more at risk than females for automobile accidents? Has the frequency of Alzheimer's disease increased in recent times? Calculating frequencies helps the epidemiologist answer questions such as these and then use the answers to better understand and control health risks.

The second concept we will explore is known as the association between the frequency of an event and various risk factors. These associations can be clues to uncovering the causes of a disease or other health-related event and then possibly lead to strategies for cure or lessening that frequency in a population. For example, French citizens drink more wine than citizens in the United States. Might this explain the difference in frequency of heart disease? Do males drive more than females, and might this explain the difference in frequencies of automobile accidents? And are we seeing more people afflicted with Alzheimer's disease because of some change in the environment, or could it be that more people are afflicted simply because people are living longer?

Once we learn how to calculate the measures of frequency and association, we will learn to interpret these numbers with data from actual epidemiologic studies. Armed with a sense of these measures, you will be able to better interpret the results of such studies as they are reported in newspapers and newsmagazines.

# Teacher's Notes on Measures in Epidemiology

## Goals and Objectives of This Module

The goal of this module is to introduce the student to some elementary ideas of measurement in epidemiology. Measurement, the translation of observations into a meaningful numeric form, is commonly taught in science classes with measures of length, time and mass. For these epidemiologic concepts, measurement is derived via expression of the concepts desired and then the development of mathematical formulas that express these ideas.

## Epidemiological Concepts to Be Covered

Measures of frequency:	Incidence and prevalence rate Incidence and prevalence rate per powers of 10
Measures of association:	Rate difference Relative risk

## Prerequisites

The mathematical prerequisites generally would be placement in first-year algebra or higher. The algebra of rational expressions would be helpful, but only in enhancing a comfort zone for the student. Rational expressions in the module are presented as formulas, and no manipulation of them is needed.

## Guide for Teachers

This module is designed to be a 2- to 3-day module for first-year algebra students and a 1- to 2-day module for second-year algebra and higher students. More important than the algebraic skill of students would be their science background and capability for understanding and discussing the scientific concepts and integrating the concepts with the mathematical formulation.

There is a certain amount of drudgery in the calculations, and it is suggested that class time be given for teams of students to perform the calculations. Students are notorious when it comes to not showing their work, but they should show the first couple of calculations in some detail, appealing to the formulas and presenting their substitutions for easy monitoring by the teacher while they are working. Once they perform the calculation processes correctly, there is no particular need for them to show their work extensively.

## Web Links

The Web links for these topics are legion. Entering incidence, prevalence and relative risk (as a phrase) in [www.google.com](http://www.google.com) generates thousands of hits. Many of these are instructional materials from medical schools and health-related professional sites. Some sites offer more examples of the concepts and even have automatic calculations of relative risk, given the  $2 \times 2$  table.

If a class has easy access to a lab, the students will find it interesting to explore the risks and prevalence for various health events, e.g., alligator attacks and New York City animal bites ([www.flmnh.ufl.edu/fish/Sharks/attacks/relrisk.htm](http://www.flmnh.ufl.edu/fish/Sharks/attacks/relrisk.htm)).

## Relation to Standards of the National Council of Teachers of Mathematics

This module reinforces three NCTM standards: measurement, connections and communication. The measurement in this module takes students beyond the typical measures, connecting the concept of measurement with scientific applications. The measures of incidence and prevalence are units that function as extensions of the ideas of rate measurement, and relative risk is an example of a dimensionless measure. The connections and communication strands are represented also. Reading and writing about mathematics in the context of another discipline helps students develop a view of mathematics as an integrated whole, rather than simply abstract manipulations of symbols.

### References

An excellent and very readable account of epidemiology is

Stolley P, Lasky T. *Investigating Disease Patterns: The Science of Epidemiology*. New York: WH Freeman; 1998.

A more mathematical but still readable treatment of these ideas is

Gordis L. *Epidemiology*. 2nd ed. Philadelphia: WB Saunders; 2000.

Data for the problems mentioned in this module were taken from the following sources:

Aldoon WH, Giovannucci EL, Stampfer MJ, Rimm EB, Wing AL, Willett WC. A prospective study of alcohol, smoking, caffeine, and the risk of duodenal ulcer in men. *Epidemiology*. 1997; 8(4): 420–424.

Bertrand, J, Mars A, Boyle C, Bove F, Yeargin-Allsopp M, Decoufle P. Prevalence of autism in a United States population: the Brick Township, New Jersey, investigation. *Pediatrics*. 2001; 108(5): 1155–1161.

Greenwood M. *Epidemics and Crowd-Diseases*. New York: Arno Press; 1977.

Li D, Odouli R, Wi S, et al. A population-based prospective cohort study of personal exposure to magnetic fields during pregnancy and the risk of miscarriage. *Epidemiology*. 2002; 13(1): 9–20.

McConnell R, Berhane K, Gilliland F, et al. Indoor risk of factors for asthma in a prospective study of adolescents. *Epidemiology*. 2002; 13(3): 288–295.

# Lesson Plan

**TITLE:** Measures in Epidemiology

**SUBJECT AREA:** Science, mathematics

**TOPIC:** Biology; making connections in mathematics

**OBJECTIVES:** The student should acquire an elementary knowledge of how observations are translated into measurements that will help in understanding scientific phenomena. In this module the measures of interest are an amount of a health problem existing in a population at a point in time or over an interval and then how the amount of a health problem might be related to potential causes or risk factors.

**TIME FRAME:** 2–3 days for a quick first-year algebra or second-year algebra class.

**PREREQUISITE KNOWLEDGE:** The prerequisite science knowledge is a general knowledge of science and health; the mathematical knowledge required is an elementary capability to use formulas to calculate quantities presented as fractions. Mathematical maturity equivalent to that needed for Algebra I is suggested.

**MATERIALS NEEDED:** A calculator would greatly facilitate calculations and cut down on the time needed for the students to finish the assignments.

**PROCEDURE:** Depending on the preference of the instructor and the maturity of the students, the module could be presented in class in a lecture format or given to the students to read before class discussion. In either case, both measures of frequency and measures of association should take only 1 day each, with 1 day for a summative evaluation.

**ASSESSMENT:** Formative evaluation should be via class discussion of the two worksheets; summative evaluation via the quiz provided.

## LINK TO STANDARDS:

### Mathematics:

Elementary operations, formula calculation

Communication and connections

### Science:

Evidence, models and explanation

Nature of scientific knowledge

Available at: <http://standards.nctm.org>

**Bibliography**

Gordis L. *Epidemiology*. 2nd ed. Philadelphia: WB Saunders; 2000.

Lilienfeld DE, Stolley PD. *Foundations of Epidemiology*. New York: Oxford University Press; 1994.

Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 1998.



# Part I: Measures of Frequency

To begin, let us consider a potentially serious health problem, perhaps the most common chronic childhood disease: asthma. In a recent study more than 3,500 children and young adults with no history of asthma were identified in schools in 12 communities in Southern California. Table 1 is constructed from information in this study. For each year of the study, the young people are categorized as newly diagnosed with asthma or still free of asthma.

*Table 1. Asthma Cases: Raw Data*

---

Year	Newly Diagnosed with Asthma	No Asthma Diagnosed
0	0	3,535
1	55	3,480
2	50	3,430
3	55	3,375
4	50	3,325
5	55	3,270

---

One measure of frequency we might use is a simple count: How many students have been diagnosed with asthma at a particular time? However, the simple count has a slight flaw that makes comparison with other populations difficult—250 cases of asthma out of a population of 3500 does not have the same meaning as 250 cases out of a population of 350! To increase the interpretability of health-related data like we have, we will take advantage of fractions, because we are usually talking about events that affect only a part of the whole population.

In studying the course of a disease or other health-related event, epidemiologists are interested in two very important measures. The **prevalence** of an event refers to the total number of existing cases at a point in time. The **incidence** of a health event refers to the number of *new* cases during a certain time period. These quantities may seem very similar, but they actually serve very different purposes. The prevalence of a disease or health-related condition is useful for those in the health professions who must deliver services to the public in the form of medicine, hospital beds or medical equipment. These services depend on having enough equipment or health care providers. The incidence of a disease is used for a fundamentally different purpose. Epidemiologists want to understand what the risk factors are for a particular health problem, and they search for the sources and causes of diseases or injuries. To accomplish this goal they will need to measure different variables and assess whether or not they are related to the number of new cases of a disease or health-related event.

Incidence and prevalence are defined as fractions and are usually presented in decimal form. When fractions are defined with time as the measure in the denominator, they are usually referred to as rates, and therefore the terms **incidence rate** and **prevalence rate\*** are defined as follows:

$$\text{Incidence rate} = \frac{\left( \begin{array}{l} \text{number of new cases of a disease occurring in} \\ \text{the population during a specified period time} \end{array} \right)}{\left( \begin{array}{l} \text{number of persons exposed to risk of developing} \\ \text{the disease during that period of time} \end{array} \right)}$$

$$\text{Prevalence rate} = \frac{\left( \begin{array}{l} \text{number of cases of disease present in} \\ \text{the population at a specified time} \end{array} \right)}{\left( \begin{array}{l} \text{number of persons at risk of having} \\ \text{the disease at that specified time} \end{array} \right)}$$

The good news about health-related events is that they are fairly rare. The bad news is that the rarity of these events can lead to pretty small decimals that are difficult to interpret easily because there are lots of zeros before the first significant digit. For this reason, the incidence and prevalence rates are frequently expressed in the form of cases per 100 or cases per 1000 for easier understanding. For example, incidence and prevalence rates per 1,000 might be expressed using the following formula:

$$\text{Incidence rate per 1000} = \frac{\left( \begin{array}{l} \text{number of new cases of a disease occurring in} \\ \text{the population during a specified period of time} \end{array} \right)}{\left( \begin{array}{l} \text{number of persons exposed to risk of developing} \\ \text{the disease during that period of time} \end{array} \right)} \times 1,000$$

$$\text{Prevalence rate per 1000} = \frac{\left( \begin{array}{l} \text{number of cases of disease present in} \\ \text{the population at a specified time} \end{array} \right)}{\left( \begin{array}{l} \text{number of persons at risk of having} \\ \text{the disease at that specified time} \end{array} \right)} \times 1,000$$

Using data for the sample of Californians above, we calculate the incidence and prevalence rates for the years 1 through 5. The numbers needed for the formulas are presented in the table. As you read the table, note that at the end of year 1 there were 55 cases of asthma diagnosed, leaving 3,480 individuals at risk of getting asthma. In year 2, there was an increase of 50 cases.

\*Strictly speaking, prevalence is a proportion, not a rate. However, "prevalence rate" is a term frequently used.

*Table 2a. Incidence and Prevalence Calculations: Fractions*

<b>Year</b>	<b>New Cases</b>	<b>Total Cases So Far</b>	<b>Individuals Still at Risk</b>	<b>Annual Incidence Rate</b>	<b>Prevalence Rate</b>
0	0	0	3,535	—	0/3,535
1	55	55	3,480	55/3,535	55/3,535
2	50	105	3,430	50/3,480	105/3,535
3	55	160	3,375	55/3,430	160/3,535
4	50	210	3,325	50/3,375	210/3,535
5	55	265	3,270	55/3,325	265/3,535

Converting these fractions to decimals gives the following results:

*Table 2b. Incidence and Prevalence Calculations: Decimals*

<b>Year</b>	<b>New Cases</b>	<b>Total Cases So Far</b>	<b>Individuals Still at Risk</b>	<b>Annual Incidence Rate</b>	<b>Prevalence Rate</b>
1	55	55	3,480	0.0156	0.0156
2	50	105	3,430	0.0144	0.0297
3	55	160	3,375	0.0160	0.0453
4	50	210	3,325	0.0148	0.0594
5	55	265	3,270	0.0165	0.0750

Notice that our incidence and prevalence rates are decimals with the first significant digit in the hundredth place. As mentioned previously, from a psychologic standpoint these are not usually easily interpreted. These rates are expressed per 100 or per 1,000 to make interpretation easier; because we have thousands of individuals in our sample, it would make some sense to express these decimals per 1,000.

*Table 2c. Incidence and Prevalence Calculations: Rates per 1,000*

---

<b>Year</b>	<b>Annual Incidence Rate per 1000</b>	<b>Prevalence Rate per 1000</b>
1	15.6	15.6
2	14.4	29.7
3	16.0	45.3
4	14.8	59.4
5	16.5	75.0

---

In this part of the module, we have developed some of the terminology and methods of calculation that epidemiologists use to describe the occurrence of a disease or other health-related event. As you can see, the terminology signifies particular concepts that can be expressed mathematically in a formula form. We will now give you a chance to practice with these concepts and formulas in Worksheet 1.

# Worksheet 1:

## Practice Problems in Prevalence and Incidence (Student Version)

Name \_\_\_\_\_

*Directions:* Using the definitions of incidence and prevalence given previously, answer the following questions. Be sure to show your work.

1. Spring fever! Epidemiologists and teachers alike have noticed the cyclic occurrence of the dreaded spring fever. Students afflicted with this disease exhibit certain listlessness in class, and they seem to stare out the windows with remarkable tenacity. Why, it's almost as if the trees growing outside are more interesting than doing algebra. (Gasp!)

At Metropolitan High School, teachers have been monitoring the situation and diagnosing spring fever. Their observations began with week 0, the first week in May. The teachers' data are shown in the following table:

### *Spring Fever Cases: Raw Data*

---

<b>Week</b>	<b>Newly Diagnosed with SF</b>	<b>No SF Diagnosed</b>
0	0	1,025
1	105	920
2	180	740
3	390	350
4	325	25

---

- a. For weeks 1–4, calculate the prevalence and incidence rates, and express them as decimals.

## *Spring Fever: Incidence and Prevalence rates*

---

<b>Week</b>	<b>New Cases</b>	<b>Total Cases So Far</b>	<b>Individuals at Risk at Start of Week</b>	<b>Weekly Incidence Rate</b>	<b>Prevalence Rate</b>
1					
2					
3					
4					

- b. From your data above, convert the incidence and prevalence rates to rates per 100 and fill in the table below with your answers. (Note that week 1 is the first row of changes where calculations should be made.)

## *Spring Fever: Incidence and Prevalence Rates per 100*

---

<b>Week</b>	<b>Weekly Incidence Rate</b>	<b>Prevalence Rate</b>
1		
2		
3		
4		

- c. Consider the incidence rates you have calculated. Based on these data, when should teachers expect the greatest increases in spring fever?
2. Diphtheria. Diphtheria is a disease caused by bacteria, and it usually affects the tonsils, throat, nose and/or skin. It is passed from person to person via coughing or sneezing, but it can also be spread by drinking from a glass used by an infected person. In 1878 Sir William H. Powers was investigating an outbreak of diphtheria in Kilburn and St. Johns Wood in England. As part of his early epidemiologic detective work, Powers noticed that if he drew two concentric circles at a particular point—one with a radius of a half-mile, one with a radius of one mile—an interesting pattern to the disease emerged. He studied the data for the inner circle and the outer ring. Here are his data for 15 weeks of observation:

## *The Incidence of Diphtheria: 1878 England*

---

<b>Week Ending</b>	<b>Inner Circle: Number of Households Newly Attacked</b>	<b>Outer Ring: Number of Households Newly Attacked</b>
March 9	1	1
March 16	4	1
March 23	2	0
March 30	4	1
April 6	0	2
April 13	0	2
April 20	2	0
April 27	5	1
May 4	8	2
May 11	7	0
May 18	41	5
May 25	14	1
June 1	5	3
June 8	3	0
June 15	2	1

- a. There were 700 houses in Powers's inner circle and 2,700 in the outer circle (for a total of 3,400). Using the end of the week of March 2nd as the "zero time," calculate the incidence rates for these 15 weeks in the inner and outer circle. Show your results in the table below.

*Incidence Rates for the Circles*

Week Ending	Inner Circle: Number of Households		Inner Circle: Number of Households Still at risk		Inner Circle: Incidence Rate		Outer Ring: Number of Households Newly Attacked		Outer Ring: Number of Households Still at Risk		Outer Ring: Incidence Rate	
	Newly Attacked											
March 2	0											
March 9												
March 16												
March 23												
March 30												
April 6												
April 13												
April 20												
April 27												
May 4												
May 11												
May 18												
May 25												
June 1												
June 8												
June 15												



- b. Make two graphs from your data. (The week ending March 2 = 0.)
  - i. A line plot of the incidence rate per week for the inner circle versus the week number.
  - ii. A line plot of the incidence rate per week for the outer ring versus the week number.

From your graphs, how would you describe the pattern of incidence of diphtheria?

3. Autism. Autism is a serious and lifelong disability that is characterized by a severely decreased ability to engage in communication and social interaction. In 1998 citizens in a New Jersey town were concerned about the number of children diagnosed with autism, and a study was undertaken to establish the prevalence in the community. The citizens were concerned about possible environmental factors that might be causing the autism, as well as establishing an appropriate level of educational and social services for the autistic children.

Data from this study are reported below:

### *Numbers of Children Diagnosed with Autistic Disorder*

---

<b>Age Category (y)</b>	<b>Diagnosed with Autistic Disorder</b>	<b>Number of Children in Population</b>
3–5	19	3,479
6–10	17	5,417

- a. Calculate the prevalence rate of autism for these children for the two age categories.
- b. Convert the prevalence rate to a rate per 1,000.

# Worksheet 1:

## Practice Problems in Prevalence and Incidence (Teacher's Answer Key)

Name \_\_\_\_\_

*Directions:* Using the definitions of incidence and prevalence given previously, answer the following questions. Be sure to show your work.

1. Spring fever! Epidemiologists and teachers alike have noticed the cyclic occurrence of the dreaded spring fever. Students afflicted with this disease exhibit certain listlessness in class, and they seem to stare out the windows with remarkable tenacity. Why, it's almost as if the trees growing outside are more interesting than doing algebra. (Gasp!)

At Metropolitan High School, teachers have been monitoring the situation and diagnosing spring fever. Their observations began with week 0, the first week in May. The teachers' data are shown in the following table:

### *Spring Fever Cases: Raw Data*

---

<b>Week</b>	<b>Newly Diagnosed with SF</b>	<b>No SF Diagnosed</b>
0	0	1,025
1	105	920
2	180	740
3	390	350
4	325	25

---

- a. For weeks 1–4, calculate the prevalence and incidence rates, and express them as decimals.

## *Spring Fever: Incidence and Prevalence Rates*

Week	New Cases	Total Cases So Far	Individuals at Risk at Start of Week	Weekly Incidence Rate	Prevalence Rate
1	105	105	1,025	0.102	0.102
2	180	285	920	0.196	0.278
3	390	675	740	0.527	0.659
4	325	1,000	350	0.929	0.976

- b. From your data above, convert the incidence and prevalence rates to rates per 100 and fill in the table below with your answers. (Note that week 1 is the first row of changes where calculations should be made.)

## *Spring Fever: Incidence and Prevalence Rates per 100*

Week	Weekly Incidence Rate per 100	Prevalence Rate per 100
1	10.2	10.2
2	19.6	27.8
3	52.7	65.9
4	92.9	97.6

- c. Consider the incidence rates you have calculated. Based on these data, when should teachers expect the greatest increases in spring fever?

**Week 4 has the highest incidence rate.**

2. Diphtheria. Diphtheria is a disease caused by bacteria, and it usually affects the tonsils, throat, nose and/or skin. It is passed from person to person via coughing or sneezing, but it can also be spread by drinking from a glass used by an infected person. In 1878 Sir William H. Powers was investigating an outbreak of diphtheria in Kilburn and St. Johns Wood in England. As part of his early epidemiologic detective work, Powers noticed that if he drew two concentric circles at a particular point—one with a radius of a half-mile, one with a radius of one mile—an interesting pattern to the disease emerged. He studied the data for the inner circle and the outer ring. Here are his data for 15 weeks of observation:

## *The Incidence of Diphtheria: 1878 England*

---

<b>Week Ending</b>	<b>Inner Circle: Number of Households Newly Attacked</b>	<b>Outer Ring: Number of Households Newly Attacked</b>
March 9	1	1
March 16	4	1
March 23	2	0
March 30	4	1
April 6	0	2
April 13	0	2
April 20	2	0
April 27	5	1
May 4	8	2
May 11	7	0
May 18	41	5
May 25	14	1
June 1	5	3
June 8	3	0
June 15	2	1

- a. There were 700 houses in Powers's inner circle and 2,700 in the outer circle (for a total of 3,400). Using the end of the week of March 2 as zero time, calculate the incidence rates for these 15 weeks in the inner and outer circles. Show your results in the table below.

*Incidence Rates for the Circles*

<b>Week Ending</b>	<b>Inner Circle: Number of Households Newly Attacked</b>	<b>Inner Circle Number of Households Still at Risk</b>	<b>Inner Circle: Incidence Rate</b>	<b>Outer Ring: Number of Households Newly Attacked</b>	<b>Outer Ring: Number of Households Still at Risk</b>	<b>Outer Ring: Incidence Rate</b>
March 2	0	700	—	0	2,700	—
March 9	1	699	1/700	1	2,699	1/2,700
March 16	4	695	4/699	1	2,698	1/2,699
March 23						
March 30						
April 6						
April 13						
April 20						
April 27						
May 4						
May 11						
May 18						
May 25						
June 1						
June 8						
June 15						

- b. Make two graphs from your data. (The week ending March 2 = 0.)
  - i. A line plot of the incidence rate per week for the inner circle versus the week number.
  - ii. A line plot of the incidence rate per week for the outer ring versus the week number.

From your graphs, how would you describe the pattern of incidence of diphtheria?

**There is a spike of incidence in mid-May, suggesting a specific cause rather than an environmental factor. (It turns out that the milk supply from a particular dairy was the culprit!)**

3. Autism. Autism is a serious and lifelong disability that is characterized by a severely decreased ability to engage in communication and social interaction. In 1998 citizens in a New Jersey town were concerned about the number of children diagnosed with autism, and a study was undertaken to establish the prevalence in the community. The citizens were concerned about possible environmental factors that might be causing the autism, as well as establishing an appropriate level of educational and social services for the autistic children.

Data from this study are reported below:

### *Numbers of Children Diagnosed with Autistic Disorder*

---

<b>Age Category (y)</b>	<b>Diagnosed with Autistic Disorder</b>	<b>Number of Children in Population</b>
3–5	19	3,479
6–10	17	5,417

- a. Calculate the prevalence rate of autism for these children for the two age categories.  
**For ages 3–5: 0.00546**  
**For ages 6–10: 0.00313**
- b. Convert the prevalence rate to a rate per 1,000.  
**For ages 3–5: 5.46 per thousand**  
**For ages 6–10: 3.13 per thousand**

## Part II: Measures of Association

Thus far we have constructed measures of the frequency of a disease or health-related event in a population. Specifically we defined and calculated the prevalence rates and incidence rates. It is frequently of interest to epidemiologists to compare these quantities for two different populations. For example, is there a greater frequency of heart disease in the population of smokers than nonsmokers? If so, it might be inferred that smoking is a risk factor for heart disease.

As another example, are individuals who have been vaccinated with vaccine A observed to have influenza less frequently than those who are not vaccinated? If so, it might be inferred that vaccine A is effective against the influenza.

If we observe that two variables are related in some way, we refer to them as being associated. In the vaccination example, one variable is the vaccination status and the other is the flu status. An individual can either have been vaccinated or not and subsequently either develop influenza or not. It is common to present information about these variables in a table form. The row and column totals are known as marginals because they appear in the margins of the table.

*Table 3a. Example of a  $2 \times 2$  Table with Marginals*

---

	<b>Influenza</b>	<b>No Influenza</b>	<b>Total</b>
<b>Not Vaccinated</b>	a	b	a + b
<b>Vaccinated</b>	c	d	c + d
<b>Total</b>	a + c	b + d	a + b + c + d

In this setup, the labels for one variable are inserted in the top row, and the labels for the other variable are inserted in the leftmost column. When a population is sampled for an epidemiologic study, an individual is observed to be in one of the four groups defined by the table: vaccinated and getting the flu (a), vaccinated and not getting the flu (b), not vaccinated and getting the flu (c), and not vaccinated and not getting the flu (d). If there is no relationship between being vaccinated and getting the flu, the table might look like this:

*Table 3b. 2 × 2 Table Illustrating No Association*


---

	Influenza	No Influenza	Total
Not Vaccinated	30	60	90
Vaccinated	20	40	60
Total	50	100	150

In this table, we can see that the incidence rate of disease for those who have been vaccinated is the same as the rate of disease for those who have not been vaccinated.

If getting vaccinated is associated with getting the flu, we might see a table that looks a bit different, possibly like Table 3c. (Notice that the exposure is to *not* being vaccinated.)

*Table 3c. 2 × 2 Table Illustrating Association*


---

	Influenza	No Influenza	Total
Not Vaccinated	35	5	40
Vaccinated	10	70	80
Total	45	75	120

The numbers in this table show that the proportion of people who were vaccinated and who subsequently got the flu is less than the proportion of people who were not vaccinated who subsequently got the flu. This sort of difference would be regarded as evidence for an association, or relationship, between vaccination and influenza.

Our task at present is to somehow quantify, or measure, this association. Fortunately we can build on our understanding of measures of frequency of a disease, specifically the incidence of a disease. We conceptually separate a population into two subpopulations: those who have been vaccinated and those who have not been vaccinated. We then observe who gets the flu; that is, we observe new cases and calculate the incidence of the flu in the two subpopulations. Our understanding of association in this example is this: If the incidence of the flu in the unvaccinated subpopulation is different from the incidence of the flu in the vaccinated subpopulation, vaccination and getting the flu are observed to be associated. The term **association** was invented by statisticians to indicate such a difference between subpopulations, but the association does not automatically imply that the two variables are related causally. It may in fact be the case that getting vaccinated causes one to be less likely to get the flu, but the mere observation of association is not enough to make the case. It could be, for example, that the people



who were vaccinated avoided getting the flu not because of the vaccination, but for other reasons related to being vaccinated. Perhaps they are wealthier; thus they can afford the vaccination and in addition do not frequent places with crowds of flu-spreading people.

Before we advance to an actual measure of association, it will be convenient to calculate the incidence rate of influenza for the two subpopulations in Table 3. Using Table 3a, we can calculate the incidence rate of the flu for those who are vaccinated and not vaccinated. It sounds a bit odd, but the exposure turns out to be exposure to *not* being vaccinated in the formula.

The incidence rate for those who have not been vaccinated and are thus exposed to risk is calculated like this:

$$\begin{aligned} \text{Incidence rate} &= \frac{\left( \begin{array}{l} \text{number of new cases of a disease occurring in} \\ \text{the population during a specified period of time} \end{array} \right)}{\left( \begin{array}{l} \text{number of persons exposed to risk of developing} \\ \text{the disease during that period of time} \end{array} \right)} \\ &= \frac{a}{a + b} = \frac{35}{35 + 5} = 0.875 \end{aligned}$$

We also calculate the incidence rate for those who have been vaccinated:

$$\begin{aligned} \text{Incidence rate} &= \frac{\text{number of new cases over a period of time}}{\text{number in the population at risk during the period}} \\ &= \frac{\text{number of new cases in the vaccinated group}}{\text{number of people vaccinated}} \\ &= \frac{c}{c + d} = \frac{10}{10 + 70} = 0.125 \end{aligned}$$

There are two elementary mathematical methods we could use to compare the incidence of the flu in these two populations, and epidemiologists use both. The first method is found by subtraction and is known as the **rate difference** (RD). In equation form,

$$\begin{aligned} \text{RD} &= \text{incidence rate among exposed} - \text{incidence rate among nonexposed} \\ &= \text{incidence rate for nonvaccinated} - \text{incidence rate for vaccinated} \\ &= \frac{a}{a + b} - \frac{c}{c + d} \\ &= 0.875 - 0.125 = 0.750 \end{aligned}$$

If vaccination and getting the flu were not associated, we would expect the two incidences to be pretty close, and thus the difference between the incidence rates would be pretty close to zero. A rate difference different from zero indicates an association, but usually the subtraction is performed so that the rate difference is positive. The rate difference is a very nice measure in this example, in that it gives a good guess about how much the incidence rate of a disease might be lessened by vaccination—assuming, of course, that a causal relation was established. On the basis of these data, we could expect to reduce the incidence of the flu from 87.5% to 12.5% through vaccination.

The second method of measuring association is more popular among epidemiologists and takes a slightly different tack than the rate difference idea. Rather than subtracting the incidences, they are divided, giving what is known as the **relative risk** (RR). With the rate difference, a difference of close to 0.0—that is, incidence rates close to each other for the exposed and nonexposed populations—indicates a lack of association between getting a disease and being exposed. With the relative risk, a quotient close to 1.0 would indicate a lack of association. The reason the relative risk is more popular than the rate difference is not obvious—it turns out that for advanced statistical procedures and more complicated situations, the relative risk is mathematically easier to use than the rate difference. It is also true that the rate difference can mask important differences in differences, to coin a phrase. The rate difference from risks of 95% and 90% and the rate difference from risks of 10% and 5% are both 5%, but their similarity masks distinctions that are very important to epidemiologists. The relative risk is defined as follows:

$$\begin{aligned} \text{RR} &= \frac{\text{incidence rate in exposed group}}{\text{incidence rate in unexposed group}} \\ &= \frac{\text{incidence rate for nonvaccinated}}{\text{incidence rate for vaccinated}} \\ &= \frac{\left( \frac{a}{a + b} \right)}{\left( \frac{c}{c + d} \right)} \end{aligned}$$

For our influenza data, the calculations would be

$$\text{RR} = \frac{\left( \frac{a}{a + b} \right)}{\left( \frac{c}{c + d} \right)} = \frac{\left( \frac{35}{35 + 5} \right)}{\left( \frac{10}{10 + 70} \right)} = 7.0$$

The relative risk is also easily interpreted with reference to possible risk factors. A relative risk close to 1.0 would indicate that the incidences for the exposed and nonexposed populations are about the same. A relative risk greater than 1.0 would indicate that exposure may well be a risk factor, and a relative risk less than 1.0 would indicate exposure to a factor that seems to decrease the risk of—in our example—getting the flu. Just how far greater or less than 1.0 a relative risk must be is difficult to specify and depends on the actual number of people involved in the study. For our data, the risk factor of not being vaccinated appears to be very large. A person who is not vaccinated is seven times as likely to develop the flu (no wonder your parents were insistent on getting those flu shots).

To illustrate these concepts with an actual example, we shall consider a study of the health effect of magnetic fields (MFs) of extremely low frequency, conducted in the San Francisco area from October 1996 through October 1998.

The investigators in this study were interested in the potential effects of magnetic fields on human pregnancies. Specifically they were interested in miscarriage. The investigators interviewed pregnant women and asked them to wear a device that would measure their exposure to magnetic field strength for 24 hours and also to keep a diary of their activities. The investigators did not observe an association between the risk of miscarriage and the average level of magnetic fields but did find an association between high magnetic field levels and the incidence of miscarriage. After an initial look at their data, they focused on the relationship between whether or not the women were exposed to magnetic fields above the level of 16 milligauss and the subsequent incidence of miscarriage. The following table presents the data on the maximum magnetic field versus whether or not the women miscarried.

### *Daily Maximum Exposure During Pregnancy vs. Miscarriage*

	Miscarriage	No Miscarriage	Total
<b>Max MF &gt; 16 mG</b>	132	585	717
<b>Max MF ≤ 16 mG</b>	27	225	252
<b>Total</b>	159	810	969

We begin our calculations by determining the rate difference of miscarriages due to exposure to high levels of magnetic fields.

$$\begin{aligned}
 \text{RD} &= \text{incidence rate among exposed} - \text{incidence rate among nonexposed} \\
 &= \text{incidence rate for MFs} > 16 \text{ mG} - \text{incidence rate for MFs} \leq 16 \text{ mG} \\
 &= \frac{a}{a + b} - \frac{c}{c + d} \\
 &= \frac{132}{132 + 585} - \frac{27}{27 + 225} \\
 &= 0.184 - 0.107 \\
 &= 0.077
 \end{aligned}$$

On the basis of this calculation, we might expect that the incidence of miscarriages could be reduced from 0.184 to about 0.107 if women could avoid contact with high magnetic fields. Again, remember that this expectation is based on the (undemonstrated) assumption that the relation between high magnetic fields and miscarriages is a causal one.

The relative risk of high magnetic fields is calculated as follows:

$$\begin{aligned}
 \text{RR} &= \frac{\text{incidence rate in exposed group}}{\text{incidence rate in unexposed group}} \\
 &= \frac{\text{incidence rate for high MF group}}{\text{incidence rate for low MF group}} \\
 &= \frac{\left( \frac{a}{a + b} \right)}{\left( \frac{c}{c + d} \right)}
 \end{aligned}$$

For the data on magnetic fields, the relative risk calculations would be

$$\text{RR} = \frac{\left( \frac{a}{a + b} \right)}{\left( \frac{c}{c + d} \right)} = \frac{\left( \frac{132}{132 + 585} \right)}{\left( \frac{27}{27 + 225} \right)} = 1.72$$

This value indicates that a woman exposed to high magnetic fields is about 1.7 times as likely to miscarry as a woman not exposed.

In this part of the module, we built on our measures of frequency and developed some of the methods used by epidemiologists to express the association between risk factors and the onset of a disease or other health-related event. These measures are fundamental tools in the epidemiologist's search for the cause of health-related problems and, as we have seen, can be used to quantify the comparative risk that exposure to a particular factor brings to a community. We will now give you a chance to practice with these concepts and formulas in Worksheet 2.

# Worksheet 2: Practice Problems in Measures of Association (Student Version)

Name \_\_\_\_\_

1. Describe in your own words what it means for two variables to be associated.
2. Suppose it were true that studying was a risk factor for better grades. To specify these variables, we will offer these values: student could study less or more and could get better or worse grades. Fill in the top row and left column of the table below with variable names that would lead to a calculation of an association between these variables. (Hint: Put the risk factor in the left column, and the outcomes in the top row.)

### *Grades vs. Studying: Round 1*

---

		<b>Total</b>
<b>Total</b>		

3. Suppose that of 80 students who studied more, 50 of them got better grades, and of 60 students who studied less, 35 got better grades. Fill in the table below and calculate the rate difference and the relative risk. Use your results to assess the risk posed by studying. Is it possible that students who study more are at risk of better grades, based on these data?

### *Grades vs. Studying: Round 2*

---

		<b>Total</b>
<b>Total</b>		

4. Smoking and duodenal ulcers. The Health Professionals Follow-up Study is a prospective study of heart disease and cancer among more than 50,000 health professionals in the United States who were 40–75 years of age in 1986. Every two years questionnaires are sent to these individuals, and newly diagnosed cases of various diseases are reported. The following data are constructed from the surveys returned in the 1992 mailing. The investigators in this study were interested in the relationship between smoking status and duodenal ulcers, a common disorder of the gastrointestinal tract. The incidence of duodenal ulcers for three groups is presented below:

### *Smoking and the Incidence of Duodenal Ulcers*

	Number of Persons at Start of Study	Number of Observed New Cases of Duodenal Ulcers
<b>Nonsmokers</b>	22,295	60
<b>Past Smokers</b>	20,757	60
<b>Current Smokers</b>	4,754	16

- a. Calculate the relative risks of being a past smoker and a current smoker, relative to never having smoked. (Hint: You should make *two* tables, one for smokers and one for past smokers.)
- b. It is sometimes said by smokers, "The damage has been done, so I might as well keep smoking." Others believe that if they quit "right now" their risk will be decreased. Which view is supported by the relative risks you calculated above?
- c. In your judgment, what values for the relative risks would support the opposite view from the one you believed was supported in part b?

# Worksheet 2: Practice Problems in Measures of Association (Teacher's Answer Key)

Name \_\_\_\_\_

1. Describe in your own words what it means for two variables to be associated.  
**Answers will vary, but generally students should say that the distribution of outcomes is different for the exposed and unexposed groups.**
  
2. Suppose it were true that studying was a risk factor for better grades. To specify these variables, we will offer these values: Student could study less or more and could get better or worse grades. Fill in the top row and left column of the table below with variable names that would lead to a calculation of an association between these variables. (Hint: Put the risk factor in the left column and the outcomes in the top row.)

### *Grades vs. Studying: Round 1*

---

	Better Grades	Worse Grades	Total
Study More			
Study Less			
Total			

3. Suppose that of 80 students who studied more, 50 of them got better grades, and of 60 students who studied less, 35 got better grades. Fill in the table below and calculate the rate difference and the relative risk. Use your results to assess the risk posed by studying. Is it possible that students who study more are at risk of better grades, based on these data?



## Grades vs. Studying: Round 2

	Better Grades	Worse Grades	Total
<b>Study More</b>	50	30	80
<b>Study Less</b>	35	25	60
<b>Total</b>	85	55	140

$$\text{Rate difference} = \frac{50}{80} - \frac{35}{60} = 0.041$$

$$RR = \frac{\frac{50}{80}}{\frac{35}{60}} = 1.07$$

Since  $1.07 > 1$ , it appears that studying does put one at risk for better grades.

4. Smoking and duodenal ulcers. The Health Professionals Follow-up Study is a prospective study of heart disease and cancer among more than 50,000 health professionals in the United States who were 40–75 years of age in 1986. Every two years questionnaires are sent to these individuals, and newly diagnosed cases of various diseases are reported. The following data are constructed from the surveys returned in the 1992 mailing. The investigators in this study were interested in the relationship between smoking status and duodenal ulcers, a common disorder of the gastrointestinal tract. The incidence of duodenal ulcers for three groups is presented below:

### Smoking and the Incidence of Duodenal Ulcers

	Number of Persons at Start of Study	Number of Observed New Cases of Duodenal Ulcers
<b>Nonsmokers</b>	22,295	60
<b>Past Smokers</b>	20,757	60
<b>Current Smokers</b>	4,754	16

- a. Calculate the relative risks of being a past smoker and a current smoker, relative to never having smoked. (Hint: You should make *two* tables, one for smokers and one for past smokers.)

*Past vs. Nonsmoker*

	New Cases	Noncases
<b>Past Smoker</b>	60	20,697
<b>Nonsmoker</b>	60	22,235

RR = 1.074

*Past vs. Current Smoker*

	New Cases	Noncases
<b>Current Smoker</b>	16	4,738
<b>Nonsmoker</b>	60	22,235

RR = 1.25

- b. It is sometimes said by smokers, "The damage has been done, so I might as well keep smoking." Others believe that if they quit "right now" their risk will be decreased. Which view is supported by the relative risks you calculated above?

**Because the relative risk is less for former smokers, the "quit right now" view is supported.**

- c. In your judgment, what values for the relative risks would support the opposite view from the one you believe was supported in part b?

**If the relative risks were the same.**



## Dental Erosion in 5-Year-Old Children

---

	Evidence of Erosion	No Evidence of Erosion	Total
<b>Fluoridated Area</b>	46	30	76
<b>Nonfluoridated Area</b>	77	37	114
<b>Total</b>	123	67	190

- a. What is the incidence rate of erosion over the five years for the two groups of children? (You may assume their teeth were free of erosion at birth.)
- b. What is the rate difference between the fluoridated areas and nonfluoridated areas?
- c. What is the relative risk for those in the nonfluoridated group? (Hint: The nonfluoridated group are the exposed children.)

- d. Based on your data above, does fluoridation appear to confer a risk of increased dental erosion or a protective tendency? Justify your response by appealing to the numeric value you calculated in part c.

# Quiz: Measures in Epidemiology (Teacher's Answer Key)

Name \_\_\_\_\_

1. We learned about two measures of frequency commonly used in epidemiology, the incidence rate and the prevalence rate. In a few sentences, tell how these two measures differ.

**The prevalence rate is the ratio of the number of cases to the population.**

**The incidence rate is the ratio of the number of *new cases in a period of time* to the remaining population at risk.**

2. In Australia on December 31, 1995, there were 4,494 dialysis patients and 4,209 transplant-dependent patients. At that time, there were approximately 9.3 million citizens in Australia.
- a. What is the prevalence rate for dialysis patients?

$$\frac{4,494}{9,300} = 4.832(10)^{-4}$$

- b. What is the prevalence rate *per million* transplant-dependent patients?

$$\frac{4,209}{9.3} = 452.6 \text{ per million}$$

3. In a recent study of dental erosion in 5-year-old children, 202 healthy school-attending children were selected for study. The investigators recorded the erosion level on their maxillary deciduous incisors and whether or not their communities used fluoridated water. The results are summarized in the table below:

## *Dental Erosion in 5-Year-Old Children*

	Evidence of Erosion	No Evidence of Erosion	Total
Fluoridated Area	46	30	76
Nonfluoridated Area	77	37	114
<b>Total</b>	123	67	190

- a. What is the incidence rate of erosion over the five years for the two groups of children?  
(You may assume their teeth were free of erosion at birth.)

$$\text{For fluoridated area: } \frac{46}{76} = 0.605$$

$$\text{For nonfluoridated area: } \frac{77}{114} = 0.675$$

- b. What is the rate difference between the fluoridated areas and not fluoridated areas?

$$0.675 - 0.605 = 0.07$$

- c. What is the relative risk for those in the nonfluoridated group? (Hint: The nonfluoridated group are the exposed children.)

$$RR = \frac{0.675}{0.605} = 1.11$$

- d. Based on your data above, does fluoridation appear to confer a risk of increased dental erosion or a protective tendency? Justify your response by appealing to the numeric value you calculated in part c.

**Fluoridation appears to offer a protective effect. The relative risk for nonfluoridation is greater than 1.**