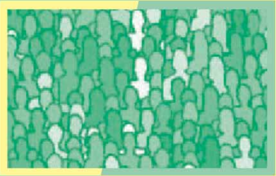
A Pocket Guide to Epidemiology



David G. Kleinbaum
Kevin M. Sullivan
Nancy D. Barker

Springer

CHAPTER 2

THE BIG PICTURE - WITH EXAMPLES

The field of epidemiology was initially concerned with providing a methodological basis for the study and control of population epidemics. Now, however, epidemiology has a much broader scope, including the study of both acute and chronic diseases, the quality of health care, and mental health problems. As the focus of epidemiologic inquiry has broadened, so has the methodology. In this overview chapter, we describe examples of epidemiologic research and introduce several important methodological issues typically considered in such research.

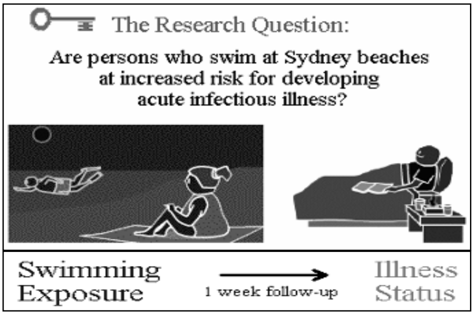
The Sydney Beach Users Study

Epidemiology *is primarily concerned with identifying the important factors or variables that influence a health outcome of interest. In the Sydney Beach Users Study, the key question was "Is swimming at the beaches in Sydney associated with an increased risk of acute infectious illness?"*

In Sydney, Australia, throughout the 1980s, complaints were expressed in the local news media that the popular public beaches surrounding the city were becoming more and more unsafe for swimming. Much of the concern focused on the suspicion that the beaches were being increasingly polluted by waste disposal.

In 1989, the New South Wales Department of Health decided to undertake a study to investigate the extent to which swimming and possible pollution at 12 popular Sydney beaches affected the public's health, particularly during the s Summer months when the beaches were most crowded. The primary research question of interest was: *are persons who swim at Sydney beaches at increased risk for developing an acute infectious illness?*

The study was carried out by selecting subjects on the beaches throughout the summer months of 1989-90. Those subjects eligible to participate at this initial interview were then followed-up by phone a week later to determine swimming exposure on the day of the beach interview and subsequent illness status during the week following the interview.


The Research Question:
Are persons who swim at Sydney beaches at increased risk for developing acute infectious illness?
Swimming Exposure
1 week follow-up
Illness Status

Water quality measurements at the beaches were also taken on each day that subjects were sampled in order to match swimming exposure to pollution levels at the beaches.

Analysis of the study data lead to the overall conclusion that swimming in polluted water carried a statistically significant 33% increased risk for an infectious illness when compared to swimming in non-polluted water. These results were considered by health department officials and the public alike to confirm that swimming in Sydney beaches posed an important health problem. Consequently, the state and local health departments together with other environmental agencies in the Sydney area undertook a program to reduce sources of pollution of beach water that lead to improved water quality at the beaches during the 1990's.

Summary

- ❖ The Sydney Beach Users Study is an example of the application of epidemiologic principles and methods to investigate a localized public health issue.
- ❖ The key question in the Sydney Beach Users Study was:
  - o Does swimming at the beaches in Sydney, Australia (in 1989-90) pose an increased health risk for acute infectious illnesses?
  - o The conclusion was yes, a 33% increased risk.

Important Methodological Issues

We provide a general perspective of epidemiologic research by highlighting several broad issues that arise during the course of most epidemiologic investigations.

There are many issues to worry about when planning an epidemiologic research study (see Box below). In this chapter we will begin to describe a list of broad methodological issues that need to be addressed. We will illustrate each issue using the previously described Sydney Beach Users Study of 1989.

Issues to consider when planning an epidemiologic research study

Question	Define a question of interest and key variables
Variables	What to measure: exposure (**E**), disease (**D**), and control (**C**) variables
Design	What study design and sampling frame?
Frequency	Measures of disease frequency
Effect	Measures of effect
Bias	Flaws in study design, collection, or analysis
Analysis	Perform appropriate analyses

The first two issues require clearly defining the study **question** of interest, followed by specifying the key **variables** to be measured. Typically, we first should ask: *What is the relationship of one or more hypothesized determinants to a disease or health outcome of interest?*

determinants
health outcome

A *determinant* is often called an **exposure variable** and is denoted by the letter **E**. The disease or health outcome is often denoted as **D**. Generally, variables other than exposure and disease that are known to predict the health outcome must be taken into account. We often call these variables **control variables** and denote them using the letter **C**.

Next, we must determine how to actually measure these variables. This step requires determining the information-gathering instruments and survey questionnaires to be obtained or developed.

The next issue is to select an appropriate **study design** and devise a sampling plan for enrolling subjects into the study. The choice of study design and sampling plan depends on feasibility and cost as well as a variety of characteristics of the population being studied and the study purpose.

Measures of disease frequency and effect then need to be chosen based on the study design. A measure of **disease frequency** provides quantitative information about how often a health outcome occurs in subgroups of interest. A **measure of effect** allows for a comparison among subgroups.

We must also consider the potential **biases** of a study. Are there any flaws in the study design, the methods of data collection, or the methods of data analysis that could lead to spurious conclusions about the exposure-disease relationship?

Finally, we must perform the appropriate **data analysis**, including stratification and mathematical modeling as appropriate. Analysis of epidemiologic data often includes taking into account other previously known risk factors for the health outcome. Failing to do this can often distort the results and lead to incorrect conclusions.

<u>Summary: Important Methodological Issues</u>

- ❖ What is the study question?
- ❖ How should the study variables be measured?
- ❖ How should the study be designed?
- ❖ What measures of disease frequency should be used?
- ❖ What kinds of bias are likely?
- ❖ How do we analyze the study data?

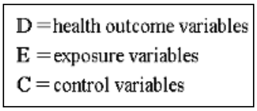
Data is obtained from:
Surveys
Interviews
Samples
Laboratories

Terms to learn
Study Designs:
Clinical trials
Cross-sectional
Case-control
Cohort
Measures of Disease Frequency
Rate
Proportion
Risk
Odds
Prevalence
Incidence
Measures of Effect
Risk ratio
Odds ratio
Rate ratio
Prevalence ratio
Biases
Selection bias
Information bias
Confounding bias
Data Analysis
Logistic Regression
Risk factors
Confounding
Effect modification

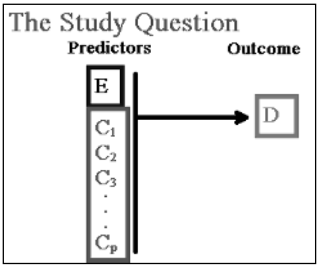
The Study Question

Epidemiology is primarily concerned with identifying the important factors or variables that influence a health outcome of interest. Therefore, an important first step in an epidemiologic research study is to carefully state the key study question of interest.

The study question needs to be stated as clearly and as early as possible, particularly to indicate the variables to be observed or measured. A typical epidemiologic research question describes the relationship between a health outcome variable, **D**, and an exposure variable, **E**, taking into account the effects of other variables already known to predict the outcome (**C,** control variables).



D = health outcome variables
E = exposure variables
C = control variables



A simple situation, which is our primary focus throughout the course, occurs when there is only one **D** and one **E**, and there are several control variables. Then, the typical research question can be expressed as shown below, where the arrow indicates that the variables **E** and the controls (**C**s) on the left are the variables to be evaluated as predictors of the outcome **D**, shown on the right.

In the Sydney Beach Users Study, the health outcome variable, **D**, of interest is whether or not a person swimming at a beach in Sydney develops an acute infectious illness such as a cough, cold, flu, ear infection, or eye infection, within one week of swimming at the beach.
The study subjects could be classified as either:

D=0 for those did not get ill, or **D**=l for those became ill.

A logical choice for the exposure variable is the exposure variable *swimming status*, which is set to:

E=0 for non-swimmers and **E**=1 for swimmers during the time period of the study.

(Note that other coding schemes could be used other than 0/1, such as 1/2, Y/N, or +/-, but we will use 0/1).

Control variables might include pollution level at the beach, age of the subject, and duration of swimming. Generally speaking, a study will not be very useful unless a question or hypothesis of some kind can be formulated to justify the time and expense needed to carry out the study.

Thus, the research question of this study example is to describe the relationship of swimming to the development of an infectious illness, while taking into account the effects of relevant control variables such as pollution level, age of subject and duration of swimming.

Because several variables are involved, we can expect that a complicated set of analyses will be required to deal with all the possible relationships among the variables involved.

Summary: The Study Question

- ❖ An important first step in an epidemiologic research study is to carefully state the key study question of interest.
- ❖ The general question: To what extent is there an association between one or more exposure variables (**E**s) and a health outcome (**D**), taking into account (i.e., controlling for) the possible influence of other important covariates (**C**s)?
- ❖ We can expect a complicated set of analyses to be required to deal with all possible relationships among the variables involved.

Quiz (Q2.1)

In the Sydney Beach Users study, exposure was alternatively defined by distinguishing those who swam in polluted water from those who swam in non-polluted water and from those who did not swim at all. Based on this scenario, fill in the missing information in the following statement:

1. The exposure variable has ??? categories, one of which is ???

Choices: **2 3 4 5 did not swim polluted water swam water not polluted**

2. When considering both swimming and pollution together, which of the following choices is appropriate for defining the exposure variable in the Sydney Beach Users study: ???

Choices:

a) E=O if did not swim, E=1 if swam in polluted water
b) E=O if did not swim, E=1 if swam in non-polluted water
c) E=O if did not swim, E=1 if swam in polluted water, E=2 if swam in non-polluted water
d) E=O if did not swim, E=1 if swam

In the Sydney Beach Users study, the illness outcome was whether or not an acute infectious illness developed 1 week after swimming at the beach. Also, in addition to age, another control variable was whether or not a study subject swam on days other than the day he or she was interviewed. Fill in the missing information:

3. The health outcome has ??? categories.
4. There are at least ??? control variables.
5. Which of the following choices is not a control variable: ???
   a) age b) swimming status on other days c) swimming status on day of interview

Choices: **2 3 4 5 a b c**

Measuring the Variables

Another important issue is: How do we measure the variables to be studied? Several measurement issues are now introduced.

Once the study question is determined, the investigators must determine how to measure the variables identified for the study and any other information that is needed. For example, how will the exposure variable be measured? If a subject went into the water but never put his head under the water, does that count as swimming? How much time is required to spend in the water to be counted as swimming? Is it feasible to observe each subject's swimming status on the day of initial interview, and if not, how should swimming status be determined?

After considering these questions, the study team defined swimming as *any immersion of the face and head in the water*. It was decided that subject self-reporting of swimming was the only feasible way to obtain swimming information.

Measuring Exposure Variables
Definition of Swimming
Any immersion of face & head in water
Measuring Swimming Status
Subject self-reporting

How will the health outcome be measured? Should illness be determined by a subject's self-report, which might be inaccurate, or by a physician's confirmation, which might not be available? The study team decided to use self-reported symptoms of illness obtained by telephone interview of study subjects 7 to 10 days after the initial interview.

Another measurement issue concerned how to determine water quality at the beach. Do water samples need to be collected? What time of day should they be collected? How will such information be linked to study subjects? The study team decided that health department surveyors would collect morning and evening samples at the midpoint of each of three sectors of the beach.

As nearly as could practicably be achieved, study subjects were to be interviewed during the period in which water samples were taken. A standard protocol was determined for how much water was to be sampled and how samples were to be assessed for water quality.

A final measurement issue concerned what information should be obtained from persons interviewed at the beach for possible inclusion into the study? The study team decided to collect basic demographic data including age, sex, and postcode, to ask whether or not each respondent had been swimming anywhere in the previous 5 days, and had any condition that precluded swimming on the day of the interview.

Interview Variables
Age
Sex
Postcode
Swimming history
Health status

Subjects were excluded from the study if they reported swimming in the previous 5 days or having an illness that prevented them from swimming. Subjects were included if they were at least 15 years old and agreed to both an initial beach interview and a follow-up telephone interview.

All the measurement issues described above must be addressed prior to data collection to ensure standardized information is collected and to provide a study that is both cost and time efficient.

Study Questions (Q2.2)

1. What other variables might you also consider as control variables in the Beach Users Study?
2. How do we decide which variables to measure as control variables?
3. Why should age be considered?
4. How would you deal with subjects who went to the beach on more than one day?

Summary: Measuring the Variables

General measurement issues:

- How to operationalize the way a measurement is carried out?
- Should self-reporting of exposure and/or health outcome be used?
- When should measurements be taken?
- How many measurements should be taken on each variable and how should several measurements be combined?
- How to link environmental measures with individual subjects?

The Study Design, including the Sampling Plan

Another important issue is: What ***study design*** *should be used and how should we select study subjects? Several study design issues are now introduced.*

There are a variety of study designs used in epidemiology. The Sydney Beach Users study employed a **cohort** design. A key feature of such a design is that subjects without the health outcome are followed-up over time to determine if they develop the outcome. Subjects were selected from 12 popular Sydney beaches over 41 sampling days. An initial interview with the study subjects took place on the beach to obtain consent to participate in the study and to obtain demographic information.

Persons were excluded from the study if they had an illness that prevented them from swimming on that day or if they had been swimming within the previous 5 days. It was not considered feasible to determine swimming exposure status of each subject on the day of initial interview. Consequently, a follow-up telephone interview was conducted 7 to 10 days later to obtain self-reported swimming exposure as well as illness status of each subject.

Study Questions (Q2.3)

1. How might you criticize the choice of using self-reported exposure and illnesses?
2. How might you criticize the decision to determine swimming status from a telephone interview conducted 7 to 10 days after being interviewed on the beach?

A complex sample survey design was used to obtain the nearly 3000 study participants. Six beaches were selected on any given day and included 2 each from the northern, eastern and southern areas of Sydney. Each beach was divided into three sectors, defined by the position of the swimming area flags erected by the lifeguards. Trained interviewers recruited subjects, starting at the center of each sector and moving in a clockwise fashion until a quota for that sector had been reached. Potential subjects had to be at least 3 meters apart.

Study Questions (Q2.4)

1. Why do you think potential subjects in a given sector of the beach were specified to be at least 3 meters apart?
2. Why is the Sydney Beach Users Study a cohort study?
3. A fixed cohort is a group of people identified at the onset of a study and then followed over time to determine if they developed the outcome. Was a fixed cohort used in the Sydney Beach Users Study? Explain.
4. A case-control design starts with subjects with and without an illness and looks back in time to determine prior exposure history for both groups. Why is the Sydney Beach Users study *not* a case-control study?
5. In a cross-sectional study, both exposure and disease status are observed at the same time that subjects are selected into the study. Why is the Sydney Beach Users study not a cross-sectional study?

Summary: Study Design

❖ Two general design issues:

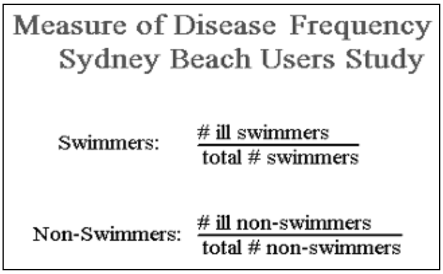
- Which of several alternative forms of **epidemiologic study designs** should be used (e.g., cohort, case-control, cross-sectional)?
- What is the **sampling plan** for selecting subjects?

Measures of Disease Frequency and Effect

*Another important issue is: What **measure of disease frequency** and **measure of effect** should be used? These terms are now briefly introduced.*

Once the study design has been determined, appropriate measures of disease frequency and effect can be specified. A measure of disease frequency provides quantitative information about how often the health outcome has occurred in a subgroup of interest.

For example, in the Sydney Beach Users Study, if we want to measure the frequency with which those who swam developed the illness of interest, we could determine the number of subjects who got ill and swam and divide by the total number who swam. The denominator represents the total number of study subjects among swimmers that had the opportunity to become ill. The numerator gives the number of study subjects among swimmers who actually became ill. Similarly, if we want to measure the frequency of illness among those who did *not* swim, we could divide the number of subjects who got ill and did not swim by the total number of non-swimming subjects.

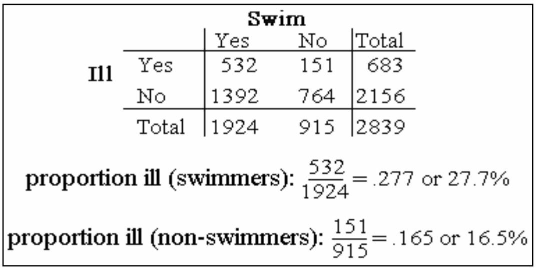


Measure of Disease Frequency
Sydney Beach Users Study

Swimmers: $\frac{\text{\# ill swimmers}}{\text{total \# swimmers}}$

Non-Swimmers: $\frac{\text{\# ill non-swimmers}}{\text{total \# non-swimmers}}$

The information required to carry out the above calculations can be described in the form of a two-way table shown below. This table shows the number who became ill among swimmers and non-swimmers. We can calculate the proportion ill among the swimmers to be 0.277 or 27.7 percent. We can also calculate the proportion ill among the non-swimmers as 0.165 or 16.5 percent.



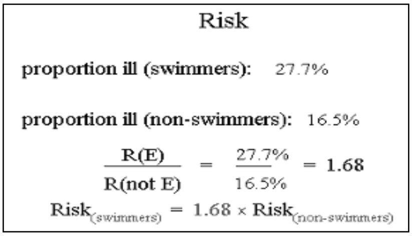
		Swim		
		Yes	No	Total
Ill	Yes	532	151	683
	No	1392	764	2156
	Total	1924	915	2839

proportion ill (swimmers): $\frac{532}{1924} = .277$ or 27.7%

proportion ill (non-swimmers): $\frac{151}{915} = .165$ or 16.5%

Each proportion is a measure of disease frequency called a **risk**. **R(E)** denotes the risk among the exposed for developing the health outcome. **R(not E)** [or **R(**$\overline{\text{E}}$**)**] denotes the risk among the ***un***exposed. There are measures of disease frequency other than risk that will be described in this course. The choice of measure (e.g., risk, odds, prevalence, or rate) primarily depends on the type of study design being used and the goal of the research study.

If we want to compare two measures of disease frequency, such as two risks, we can divide one risk by the other, say, the risk for swimmers divided by the risk for non-swimmers. We find that the ratio of these risks in our study is 1.68; this means that swimmers have a risk for the illness that is 1.68 times the risk for non-swimmers.



Risk

proportion ill (swimmers): 27.7%

proportion ill (non-swimmers): 16.5%

$$\frac{R(E)}{R(\text{not } E)} = \frac{27.7\%}{16.5\%} = 1.68$$

$$\text{Risk}_{(\text{swimmers})} = 1.68 \times \text{Risk}_{(\text{non-swimmers})}$$

Such a measure is called a **measure of effect**. In this example, the effect of interest refers to the effect of one's swimming status on becoming or not becoming ill. If we divide one risk by the other, the measure of effect or association is called a **risk ratio**. There are other measures of effect that will be described in this course (e.g., such as the risk ratio, odds ratio, prevalence ratio, rate ratio, risk difference, and rate difference). As with measures of disease frequency, the choice of effect measure depends on the type of study design and the goal of the research study.

Summary: Measures of Disease Frequency and Effect

❖ A **measure of disease frequency** quantifies how often the health outcome has occurred in a subgroup of interest.

❖ A **measure of effect** quantifies a comparison of measures of disease frequency for two or more subgroups.

❖ The choice of measure of disease frequency and measure of effect depends on the type of study design used and the goal of the research study.

Bias

Another important issue is: What are the potential biases of the study? The concept of bias is now briefly introduced.

Bias: A flaw in the
- study design
- methods of data collection
- methods of data analysis

… which leads to spurious conclusions.

Sources of bias:
- Selection
- Information
- Confounding

The next methodologic issue concerns the potential biases of a study. Bias is a flaw in the study design, the methods of data collection, or the methods of data analysis that may lead to spurious conclusions about the exposure-disease relationship. Bias may occur because of: the **selection** of study subjects; incorrect information gathered on study subjects; or failure to adjust for variables other than the exposure variable, commonly called **confounding.**

In the Sydney Beach Users Study, all 3 sources of bias were considered. For example, to avoid **selection bias**, subjects were excluded from the analysis if they were already ill on the day of the interview. This ensured that the sample represented only those healthy enough to go swimming on the day of interview. Sometimes selection bias cannot be avoided. For example, subjects had to be excluded from the study if they did not complete the follow-up interview. This

non-response bias may affect how representative the sample is.

There was also potential for **information bias** since both swimming status and illness status were based on self-reporting by study subjects. Swimming status was determined by self-report at least seven days after the swimming occurred. Also, the report of illness outcome did not involve any clinical confirmation of reported symptoms.

Confounding in the Beach Users Study concerned whether all relevant variables other than swimming status and pollution level exposures were taken into account. Included among such variables were age, sex, duration of swimming for those who swam, and whether or not a person swam on additional days after being interviewed at the beach. The primary reason for taking into account such variables was to ensure that any observed effect of swimming on illness outcome could not be explained away by these other variables.

Summary

- ❖ Bias is a flaw in the study design, the methods of data collection, or the methods of data analysis that may lead to spurious conclusions about the exposure-disease relationship.
- ❖ Three general sources of bias occur in:
  - o Selection of study subjects
  - o Incorrect information gathered on study subjects
  - o Failure to adjust for variables other than the exposure variable (confounding)

Analyzing the data

Another important issue is: How do we carry out the data analysis? We now briefly introduce some basic ideas about data analysis.

The final methodologic issue concerns the data analysis. We must carry out an appropriate analysis once collection and processing of the study data are complete. Since the data usually come from a sample of subjects, the data analysis typically requires the use of statistical procedures to account for the inherent variability in the data. In epidemiology, data analysis typically begins with the calculation and statistical assessment of simple measures of disease frequency and effect. The analysis often progresses to more advanced techniques such as stratification and mathematical modeling. These latter methods are typically used to control for one or more potential confounders.

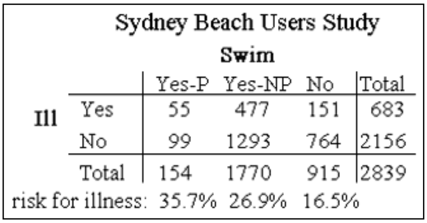
Statistics

Frequency	**Effect**
Risk	risk ratio
Prevalence	prevalence ratio
Odds	odds ratio
Rate	rate ratio

Stratification

Mathematical modeling

Let's consider the data analysis in the Sydney Beach Users Study. We had previously compared swimmers with non-swimmers. Now, we may wish to address the more specific question of whether those who swam in polluted water had a higher risk for illness than those who swam in non-polluted water. We can do this by separating the swimmers into two groups. The non-swimmers represent a baseline comparison group with which the two groups of swimmers can be compared.

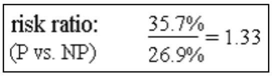


Sydney Beach Users Study

Swim

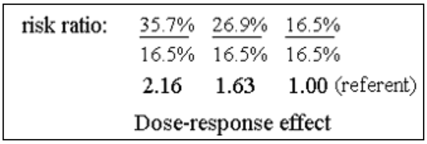
		Yes-P	Yes-NP	No	Total
Ill	Yes	55	477	151	683
	No	99	1293	764	2156
	Total	154	1770	915	2839
risk for illness:		35.7%	26.9%	16.5%	

Based on the two-way table, we can estimate the risk for illness for each of the three groups by computing the proportion that got ill out of the total for each group. The three risk estimates are 0.357, 0.269 and 0.165, which translates to 35.7 percent, 26.9 percent and 16.5 percent, respectively.



risk ratio: (P vs. NP) $\frac{35.7\%}{26.9\%} = 1.33$

The risk ratio that compares the Swam-Polluted (Yes-P) group with the Swam-Nonpolluted (Yes-NP) group is 1.33 indicating that persons who swam in polluted water had a 33 percent increased risk than persons who swam in nonpolluted water.



risk ratio:	$\frac{35.7\%}{16.5\%}$	$\frac{26.9\%}{16.5\%}$	$\frac{16.5\%}{16.5\%}$
	2.16	**1.63**	**1.00** (referent)

Dose-response effect

Also, the risk ratio estimates obtained by dividing the risks for each group by risk for non-swimmers are 2.16, 1.63, and 1. This suggests what we call a dose-response effect, which means that as exposure increases, the risk increases.

The analysis just described is called a "crude" analysis because it does not take into account the effects of other known factors that may also affect the health outcome being studied. A list of such variables might include age, swimming duration, and whether or not a person swam on additional days. The conclusions found from a crude analysis might be altered drastically after adjusting for these potentially confounding variables.

Several questions arise when considering the control of many variables:

- Which of the variables being considered should actually be controlled?
- What is gained or lost by controlling for too many or too few variables?
- What should we do if we have so many variables to control that we run out of numbers?
- What actually is involved in carrying out a stratified analysis or mathematical modeling to control for several variables?
- How do the different methods for control, such as stratification and mathematical modeling, compare to one another?

These questions will be addressed in later activities.

Study Questions (Q2.5)

1. How do you interpret the risk ratio estimate of 1.33?
2. Does the estimated risk ratio of 1.33 indicate that swimming in polluted water poses a health risk?
3. Given the relatively small number of 154 persons who swam in polluted water, what statistical question would you need to answer about the importance of the estimated risk ratio of 1.33?

Summary: Analyzing the Data

- ❖ The data analysis typically requires the use of statistical procedures to account for the inherent variability in the data.

- In epidemiology, data analysis often begins with assessment and comparison of simple measures of disease frequency and effect.
- The analysis often progresses to more advanced techniques such as stratification and mathematical modeling.

Example: Alcohol Consumption and Breast Cancer

The Harvard School of Public Health followed a cohort of about 100,000 nurses from all over the US throughout the 1980s and into the 1990s. The investigators in this Nurses Health Study, were interested in assessing the possible relationship between diet and cancer. One particular question concerned the extent to which alcohol consumption was associated with the development of breast cancer.

Nurses identified as being 'disease free' at enrollment into the study were asked about the amount of alcohol they currently drank. Other relevant factors, such as age and smoking history, were also determined. Subjects were followed for four years, at which time it was determined who developed breast cancer and who did not. A report of these findings was published in the New England Journal of Medicine in 1987.

Recall that the first methodologic issue is to define the **study question**. Which of the study questions stated here best addresses the question of interest in this study?

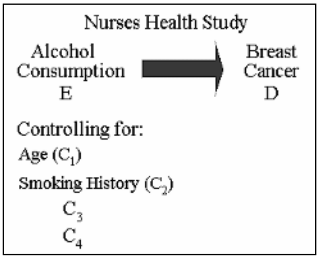
A. Is there a relationship between drinking alcohol and developing breast cancer?
B. Are alcohol consumption, age, and smoking associated with developing breast cancer?
C. Are age and smoking associated with developing breast cancer, after controlling for alcohol consumption?
D. Is alcohol consumption associated with developing breast cancer, after accounting for other variables related to the development of breast cancer?

The best answer is "D," although "A" is also correct. In stating the study question of interest, we must identify the primary variables to be measured.

<u>Study Questions (Q2.6)</u> Determine whether each of the following is a:
Health outcome variable (D), Exposure variable (E), or Control variable (C)

1. Smoking history
2. Whether or not a subject develops breast cancer during follow-up
3. Some measure of alcohol consumption
4. Age

Once we have specified the appropriate variables for the study, we must determine how to measure them. The health outcome variable, **D**, in this example is simply *yes* or *no* depending on whether or not a person was clinically diagnosed with breast cancer. The investigators at Harvard interviewed study subjects about their drinking habits, **E**, and came up with a quantitative



measurement of the amount of alcohol in units of grams per day that were consumed in an average week around the time of enrollment into the study. How to treat this variable for purposes of the analysis of the study data was an important question considered. One approach was to categorize the alcohol measurement into 'high' versus 'low'. Another approach was to categorize alcohol into 4 groups: non-drinkers; less than 5 grams per day; between 5 and 15 grams per day; and 15 or more grams per day.

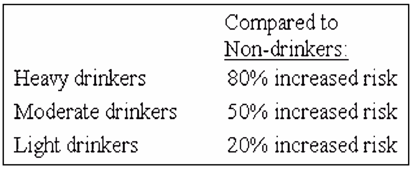
Age, denoted **C_1,** is inherently a quantitative variable, although many of the analyses treated age as a categorical variable in three age groups, shown here:

34 to 44 years, 45 to 54 years, 55 to 59 years

Smoking history, **C_2**, was categorized in several ways; one was *never* smoked versus *ever* smoked.

The research question in the nurse's health study can thus be described as determining if there is a relationship between alcohol consumption, **E**, and breast cancer, **D**, controlling for the effects of age, **C_l**, and smoking history, **C_2**, and possibly other variables (C_3, C_4, etc.).

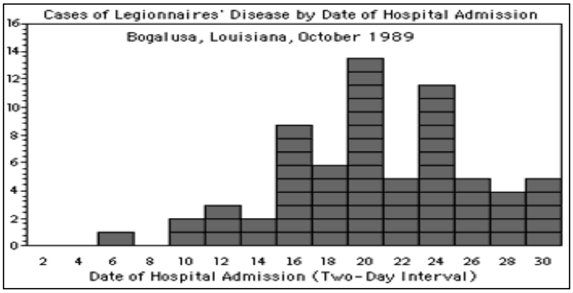
Although a detailed analysis is not described here, the data did provide evidence of a significant association between alcohol use and development of breast cancer. For heavy drinkers, when compared to non-drinkers, there was about an 80% increase in the risk of developing breast cancer. Moderate drinkers were found to have about a 50% increase in risk, and light drinkers had an increased risk of about 20%.



	Compared to Non-drinkers:
Heavy drinkers	80% increased risk
Moderate drinkers	50% increased risk
Light drinkers	20% increased risk

Note: The Nurses Health Study provides an example in which the exposure variable, alcohol consumption, has several categories rather than simply binary. Also, the control variables age and smoking history can be a mixture of different types of variables. In the Nurses Health Study, age is treated in three categories, and smoking history is treated as a binary variable.

Example: The Bogalusa Outbreak



On October 31, 1989, the Louisiana State Health Department was notified by two physicians in Bogalusa, Louisiana, that over 50 cases of acute pneumonia had occurred within a three-week interval in mid to late October, and that six persons had died. Information that the physicians had obtained from several patients suggested that the illness might have been Legionnaires Disease.

In 1989, Bogalusa was a town of about 16,000 persons. The largest employer was a paper mill located in the center of town adjacent to the main street. The paper mill included five prominent cooling towers. The mill also had three paper machines that emitted large volumes of aerosol along the main street of town. Many people suspected that the cooling towers and or the paper mill were the cause of the outbreak, since they were prominent sources of outdoor aerosols where the legionnaire's bacteria could have been located.

Recall that the first methodologic issue is to define the **study question** of interest. Which of the study questions stated here best addresses the question of interest in this study?

A. Was the paper mill the source of the Legionnaires Disease outbreak in Bogalusa?
B. What was the source of the outbreak of Legionnaires Disease in Bogalusa?
C. Why did the paper mill cause the outbreak of Legionnaires Disease in Bogalusa?
D. Was there an outbreak of Legionnaires Disease in Bogalusa?

The most appropriate study question is "B." Even though the paper mill was the suspected source, the study was not limited to that variable only, otherwise, it might have failed to collect information on the true source of the outbreak.

<u>Study Questions (Q2.7)</u> In stating the study question, we identify the primary variables to be considered in the study. Determine whether each of these variables is the health outcome variable, **D**, an exposure variable, **E**, or a control variable, **C**:

1. Exposure to the cooling towers of the paper mill?
2. Exposure to emissions of the paper machines?
3. Age of subject?
4. Visited grocery store A?
5. Visited grocery store B?
6. Diagnosed with Legionnaires Disease?
7. Visited drug store A?
8. Visited drug store B?
9. Ate at restaurant A?

The health outcome variable, **D**, indicates whether or not a study subject was clinically diagnosed with Legionnaires Disease during the three week period from mid to late October. The exposure variable is conceptually whatever variable indicates the main source of the outbreak. Since this variable is essentially unknown at the start of the study, there is a large collection of exposure variables, all of which need to be identified as part of the study design and investigated as candidates for being the primary source of the outbreak. We denote these exposure variables of interest $\mathbf{E_1}$ through $\mathbf{E_7}$. One potential control variable of interest was age, which we denoted as $\mathbf{C_1}$.

The general research question of interest in the Bogalusa outbreak was to evaluate the relationship of one or more of the exposure variables to whether or not a study subject developed Legionnaires Disease, controlling for age.

A **case-control study**, was carried out in which 28 **cases** diagnosed with confirmed Legionnaires Disease were compared with 56 non-cases or **controls**. This investigation led to the hypothesis that a misting machine for vegetables in a grocery store was the source of the outbreak. This misting machine was removed from the grocery store and sent to CDC where laboratory staff was able to isolate Legionella organisms from aerosols produced by the machine. This source was a previously unrecognized vehicle for the transmission of Legionella bacteria.

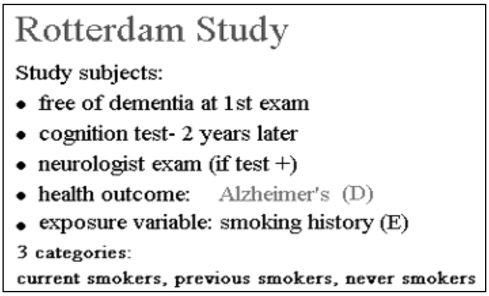
Note: The Bogalusa study provides an example in which there are several exposure variables that are candidates as the primary source of the health outcome being studied. Hopefully, the investigators will be able to identify at least one exposure variable as being implicated in the occurrence of the outbreak. It is even possible that more than one candidate exposure variable may be identified as a possible source.

The case-control study of this and many other outbreaks can often be viewed as hypothesis generating. Further study, often using laboratory methods, clinical diagnosis, and environmental survey techniques, must often be carried out in order to confirm a suspected exposure as the primary source of the outbreak. The Centers for Disease Control and Prevention has a variety of scientists to provide the different expertise and teamwork that is required, as carried out in the Bogalusa study.

Example: The Rotterdam Study

The Rotterdam study has been investigating the determinants of chronic disabling diseases, including Alzheimer's disease, during the 1990s and beyond.

In the early 1990s, the Department of Epidemiology of the Erasmus University in Rotterdam, the Netherlands, initiated the Rotterdam Study. A cohort of nearly 8000 elderly people was selected. They continue to be followed to this day. The goal of the study is to investigate determinants of chronic disabling diseases, such as Alzheimer's and cardiovascular disease. One particular study question of interest was whether smoking increases the risk of Alzheimer's disease.



Rotterdam Study

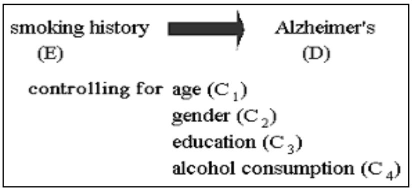
Study subjects:

- free of dementia at 1st exam
- cognition test- 2 years later
- neurologist exam (if test +)
- health outcome: Alzheimer's (D)
- exposure variable: smoking history (E)

3 categories:

current smokers, previous smokers, never smokers

Subjects who were free of dementia at a first examination were included in the study. This excluded anyone diagnosed at this exam with Alzheimer's or any other form of dementia due to organic or psychological factors. Approximately two years later, the participants were asked to take a brief cognition test. If they scored positive, they were further examined by a neurologist. The investigators could then determine whether or not a participant had developed Alzheimer's disease, the health outcome variable **D** of interest, since the start of follow-up.



smoking history (E) → Alzheimer's (D)

controlling for age (C_1)

gender (C_2)

education (C_3)

alcohol consumption (C_4)

The primary exposure variable, **E**, was smoking history. Three categories of smoking were considered: current smokers at the time of the interview; previous but not current smokers; and, never smokers. Control variables considered in this study included age, gender, education, and alcohol consumption.

We define the study question of interest as: *Is there a relationship between smoking history and Alzheimer's disease, controlling for the effects of age, gender, education and alcohol consumption?*

Recall that one important methodologic issue is to determine the study design.

How would you define the design of this study?
1. Cohort design
2. Case-control design
3. Cross-sectional design
4. Clinical trial

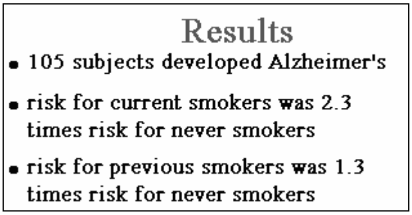
This is a cohort design because participants without the health outcome of interest, in this case Alzheimer's disease, are followed up over time to determine if they develop the outcome later in life.

Which of the following is influenced by the design of the study?

A. The assessment of confounding
B. The choice of the measures of disease frequency and effect
C. A decision regarding the use of stratified analysis
D. The analysis is not influenced in any way by the study design used

The answer is B. We determine the appropriate measures of disease frequency and effect based on the study design characteristics. Choices A and C are incorrect because they are typically considered regardless of the study design used.

The investigators found that 105 subjects developed Alzheimer's disease. After taking the control variables into account, the risk of Alzheimer's disease for current smokers was 2.3 times the risk for subjects who had never smoked. For subjects who had smoked in the past but who had given up smoking before the study started, the risk of Alzheimer's disease was 1.3 times the risk for subjects who had never smoked.



Results

- 105 subjects developed Alzheimer's
- risk for current smokers was 2.3 times risk for never smokers
- risk for previous smokers was 1.3 times risk for never smokers

<u>Study Questions (Q2.8)</u> Based on the above results:
1. What is the *percent increase* in the risk for current smokers when compared to the risk for never smokers?
2. What is the *percent increase* in the risk for previous smokers when compared to the risk for never smokers?

Because these results were statistically significant and controlled for previously established predictors of Alzheimer's, the study gave support to the hypothesis that smoking history was a significant risk factor in the development of Alzheimer's disease.

Nomenclature

C	Control variable or covariate
D	Disease or outcome variable
E	Exposure variable
R(E)	Risk among the exposed for developing the health outcome
R(not E) or **R($\bar{E}$)**	Risk among ***non***exposed for developing the health outcome
RR	Risk ratio

References

For the Sydney Beach Users Study:
Corbett SJ, Rubin GL, Curry GK, Kleinbaum DG. The health effects of swimming at Sydney Beaches. The Sydney Beach Users Study Advisory Group. Am J Public Health. 1993;83(12): 1701-6.

For the Nurses Health Study:
Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Hennekens CH, Speizer FE. Moderate alcohol consumption and the risk of breast cancer. N Engl J Med. 1987;316(19):1174-80.

For the Bogalusa Outbreak:
Mahoney FJ, Hoge CW, Farley TA, Barbaree JM, Breiman RF, Benson RF, McFarland LM. Communitywide outbreak of Legionnaires' disease associated with a grocery store mist machine. J Infect Dis. 1992;165(4):736-9.

For The Rotterdam Study:
Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol. 1991;7(4); 403-22.
Ott A, Slooter AJ, Hofman A, van Harksamp F, Witteman JC, Van Broeckhoven C, van Duijin CM, Breteler MM. Smoking and risk of dementia and Alzheimer's disease in a population-based cohort study: the Rotterdam Study. Lancet. 1998;351(9119):1840-3.

A CDC Website. The Centers for Disease control has a website called **EXCITE**, which stands for **Excellence in Curriculum Integration through Teaching Epidemiology.** The website address is http://www.cdc.gov/excite/

We suggest that you open up this website on your computer and look over the various features and purposes of the website described on the first page you see. Then click on the item (on menu on left of page) **Disease Detectives at Work** and read the first two articles entitled *Public Health on Front Burner After Sept 11* and *USA's 'Disease Detectives' Track Epidemics Worldwide*. Then click on the item (on menu on left of page) **Classroom Exercises** and go through the exercise on Legionnaires Disease in Bogalusa, Louisiana. The specific website address for this exercise is:

http://www.cdc.gov/excite/legionnaires.htm

Answers to Study Questions and Quizzes

Q2.1

1. 3, did not swim
2. C
3. 2
4. 2
5. C

Q2.2

1. General health status, smoking status, diet, including what a subject might have eaten at the beach.
2. Choose variables that are already known determinants of the health outcome. This will be discussed later under the topic of confounding.
3. Younger subjects might be less likely to get ill than older subjects.
4. In the actual study, the investigators chose to exclude subjects from the analysis if they visited the beach on days other than the day they were interviewed on the beach.

Q2.3

1. Self-reported information may be inaccurate and can therefore lead to spurious study results.

2. As with the previous question, the information obtained about exposure much later than when the actual exposure occurred may be inaccurate and can lead to spurious study results.

Q2.4

1. To minimize the inclusion in the study of a family or social groups.
2. Subjects without the health outcome, that is, healthy subjects selected at the beach, were followed-up over time to determine if they developed the outcome.
3. No, the Sydney Beach User's Study did not use a fixed cohort. Study subjects were added over the summer of 1989-90 to form the cohort.
4. Because the study started with exposed and unexposed subjects, rather than ill and not-ill subjects, and went forward rather than backwards in time to determine disease status.
5. Exposure and disease status were observed at different times for different subjects. Also, each subject was selected one week earlier than the time his or her exposure and disease status were determined.

Q2.5

1. The risk of illness for persons who swam in polluted water is estimated to be 1.33 times the risk of illness for persons who swam in non-polluted water.
2. Not necessarily. The importance of any risk ratio estimate depends on the clinical judgment of the investigators and the size of similar risk ratio estimates that have been found in previous studies.
3. Is the risk ratio of 1.33 significantly different from a risk ratio of 1? That is, could the risk ratio estimate of 1.33 have occurred by chance?

Q2.6

1. C
2. D
3. E
4. C

Q2.7

1. E
2. E
3. E
4. E
5. D
6. E
7. E
8. E

Q2.8

1. The increased risk of 2.3 translates to a 130% increase in the risk of current smokers compared to never smokers.
2. The increased risk of 1.3 translates to a 30% increase in the risk for previous smokers compared to never smokers.