Cigarette Smoking and Lung Cancer*

A *Disease Detectives* Exercise from the
Centers for Disease Control and Prevention

Teacher's Guide and Answer Key

Instructions

- Prepare students by teaching the fundamental principles of epidemiology and outbreak investigation. Materials are available on the EXCITE web site in "Background and Teaching Aids." At a minimum, students should be familiar with the basic steps of an outbreak investigation.
- One recommended format is to divide the class into small work groups of 5-10 students and have each group assign a facilitator, a recorder, and a reporter. Ask individual students from the class at large to read the narrative and questions out loud. Then have students work in their small groups to answer the questions. Finally, have the groups report their responses to the class.
- The exercise is in eleven parts. Each part should be distributed independently and only after students have completed the preceding part. The total exercise requires approximately 3¾ hours. An estimated time to completion is included for each part.
- Students may use a simple statistics program, such as Excel, for calculations.

Learning Objectives

After completing this exercise, students should be able to

- formulate a testable hypothesis;
- describe and apply the basic principles and methods of epidemiology;
- discuss the differences between case-control and cohort studies, the inferences that can be drawn from each, and the advantages and disadvantages of each;
- analyze data using odds ratios, mortality rates, rate ratios, rate differences, attributable risk percent, and population attributable risk percent;
- develop inferences and predictions based on data;
- formulate scientific explanations using logic and evidence;
- recognize alternative explanations; and
- evaluate criteria for causation.

*This exercise is adapted from a case study used each year in CDC's Epidemic Intelligence Service (EIS) Summer Course, which trains incoming EIS Officers. This adaptation was created by Dr. Natale A. Carasali while at CDC in spring 2001, fulfilling a medical epidemiology elective during his final year of training at the University of Nevada School of Medicine.

○ **1. What is the purpose of an epidemiologic study?**

Answer:

Epidemiologic studies determine whether people exposed to a certain risk factor (e.g., cigarette smoking) are significantly more likely than those not exposed to develop a certain health outcome (e.g., lung cancer)—in other words, is there a strong relationship between the exposure and the illness. Analysis is accomplished through statistical manipulation of data regarding the health problem under investigation.

○ **2. What is the hypothesis both Doll and Hill are trying to test?**

Answer:

In these studies, Hill and Doll were trying to answer the question: Is cigarette smoking linked to lung cancer? Put another way, is there a strong relationship between cigarette smoking and lung cancer.

○ **3. In the first study, why were investigators interested in interviewing patients who were hospitalized for disorders other than lung cancer?**

Teacher's Note:
A **case-control study** is used when the population under investigation is large and not well defined, as is generally the case in a communitywide outbreak. Participants are selected on the basis of the presence or absence of the ***disease outcome*** in question, and investigators ***work backward***, from effect to the suspected cause.

Answer:

These patients represent the controls, the comparison group necessary for analysis in a case-control study.

○ **4. What makes the second study a cohort study?**

Answer:

The case-subjects represent a small, well-defined population, or cohort, and they are categorized according to their exposure to a specific risk factor: cigarette smoking. In this study, the rate of death from lung cancer for smokers was compared with the rate for nonsmokers. Epidemiologists call this type of study a "prospective cohort study" because individuals were followed forward in time to determine disease occurrence.

Teacher's Note:
A **cohort study** is used when the population is small and well defined, such as in an outbreak of food poisoning among attendees of a picnic or a reception. Participants are evaluated according to their ***exposure to a suspected risk factor***, and investigators ***work forward***, from the suspected exposure to the outcome.

○ 5. Suppose you are one of Doll's investigators, and you get a call from a local hospital informing you of a newly diagnosed case of lung cancer. What is the first thing you should do?

Answer:

Verify the diagnosis. The investigator should go to the hospital and talk to the physician(s) who diagnosed the disease and look at appropriate tissue slides, x-rays, CT scans, etc. The goals in verifying the diagnosis are 1) to ensure that the illness was diagnosed properly and 2) to rule out laboratory error as the basis for the increase in diagnosed cases.

○ 6. Why is it important to try to match case-subjects and controls so closely?

Answer:

If case-subjects and controls differ only in disease status and their exposure to the variable under investigation, then disease outcome can be linked with high probability to that variable. If the two groups differ with respect to many variables, the link between a particular exposure and the disease cannot be so clearly determined.

○ 7. What are the advantages of selecting controls from the same hospitals as case-subjects?

Answer:

Controls from the same hospital are very likely to come from the same population as the case-subjects (London and surrounding communities), which should control for variables such as socioeconomic status, exposure to similar environmental conditions, access to health care, and diagnostic practices and procedures.

Other advantages include:

- Convenience
- Temporal match
- Comparable medical records keeping

- Participation (hospitalized patients are a captive audience)
- Equally heightened recall

○ 8. Why do you think Doll and his colleagues conducted their study in hospitals?

Answer:

Possibilities include

- high likelihood of finding lung cancer patients in hospitals,
- greater ability to match controls with case-subjects,
- convenient and plentiful source of controls,
- high probability of accurate diagnosis,
- availability of hospital records, and
- likelihood that patients will participate.

○ 9. Could investigators have chosen case-subjects and controls from sources besides hospitals? If so, please list some examples.

Answer:

For cases: cancer registries, death certificates, insurance files, doctors' offices, occupational records.

For controls: neighbors, friends, and acquaintances of case-subjects; patients who see the same physician but have other illnesses, a random sample of the well population.

○ 10. What population are the controls supposed to represent? How similar are they to this population?

Answer:

The controls represent age-appropriate people in London and the surrounding communities who do not have lung cancer. They are not very representative of this population, however, because hospitalized patients are generally more likely than the general population to smoke cigarettes and consume alcohol.

Teacher's Note:
If case-subjects and controls do not closely represent the intended population, the results of the study will be biased, meaning that the relationship between the exposure and the health outcome will be over- or underestimated.

○ **11. What effect will the controls have on the study's results, given the quality of their match with the population they represent?**

Answer:
The control group provides the prevalence of exposure for people without the health problem under study. Because the control population in this study tends to smoke more than the population it represents, the exposure prevalence among controls will be higher than expected. The net effect of this inconsistency is to *underestimate* the true risk of lung cancer associated with smoking.

Teacher's Note:
This is a form of selection bias because the controls who were selected are a subset, rather than a representative snapshot, of the greater population under investigation. *(For more detail on*

○ **12. From Table 1, calculate the proportion of case-subjects and controls who smoked.**

Answer:
Proportion of **case-subjects** who smoked: 1,350 / 1,357 = 99.5%
Proportion of **controls** who smoked: 1,296 / 1,357 = 95.5%

○ **13. How would you interpret these proportions?**

Answer:
The prevalence of smoking is:
- Very similar in both groups (99.5% vs. 95.5%)
- Extremely high in both groups (>95%)

○ **14. Calculate the odds of smoking for both case-subjects and controls.**

Answer:

Odds of smoking for **case-subjects**: 1350 / 7 = 192.9/1
Odds of smoking for **controls**: 1296 / 61 = 21.2/1

○ **15. Calculate the odds ratio.**

Answer:

Odds Ratio: (1350 / 7) ÷ (1296 / 61) = 192.9 ÷ 21.2 = 9.1

Teacher's Note:
The odds ratio measures the association (relationship) between an exposure and a disease. In general, the greater the value of the odds ratio, the stronger the association. In this study, the odds ratio compares the odds that a person who has lung cancer smokes cigarettes with the odds that a person who does not have lung cancer does not smoke cigarettes.

○ **16. What does this odds ratio tell us about the relationship between smoking and lung cancer?**

Answer:

The odds of being a smoker are 9.1 times higher for case-subjects (cigarette smokers) than for controls (nonsmokers). Assuming that this study is not biased, one can infer that the risk of lung cancer is *9 times higher* for cigarette smokers than for nonsmokers.

○ **17. Complete Table 2 by calculating the odds ratio for each category of cigarette consumption.**

Answer:

1–14 cigarettes:	(565/7) ÷ (706/61) = (565 x 61) ÷ (706 x 7) = 7.0
15–24 cigarettes:	(445 x 61) ÷ (408 x 7) = 9.5
25+ cigarettes:	(340 x 61) ÷ (182 x 7) = 16.3
All Smokers:	(1350 x 61) ÷ (1296 x 7) = 9.1

○ **18. How would you interpret these results?**

Answer:

The odds of acquiring lung cancer increases as the daily number of cigarettes smoked increases.

○ **19. While the study appears to demonstrate a clear association between smoking and lung cancer, cause and effect is not the only possible explanation. List other possibilities.**

Answer:

Explanations other than a true association are:

- Chance (although the data indicate that chance is unlikely)
- Selection bias
- Information bias
- Confounding
- Investigator error

Teacher's Note:
A correlation such as this, in which an increase in exposure to a risk factor increases the odds of disease, is known as a **dose-response relationship**. This relationship indicates a strong association between the risk factor and the disease.

○ **20. How do you think the 68% response rate might affect the study's results?**

Answer:

It could introduce selection bias into the study and lead to conclusions that are systematically different from the truth.

Teacher's Note:
As a general rule, we like to see response rates of 80% or better in epidemiologic studies. However, a 68% response rate is good for a survey conducted through the mail. If possible, investigators characterize the nonrespondents as best they can, to determine whether they differ from respondents on important factors.

hospitalized patients, and hospitalized patients are more likely than the general population to smoke and to consume alcohol, a selection bias exists. In this case, the bias would cause underestimation of the risk associated with smoking.

Information bias occurs when case-subjects and controls recall their histories with differing degrees of accuracy. Such bias is not highly likely in this study, for two reasons. First, since the hypothesis linking cigarette smoking with lung cancer was not widely known at the time, neither group was more likely than the other to focus primarily on their smoking histories. Second, both case-subjects and controls were hospitalized, so each person had the same chance for introspection about previous exposures or events.

Confounding can occur if a variable is 1) associated with, but not a consequence of, an exposure and 2) associated with the outcome independent of its association with the exposure. Age is a possible confounder in this study because it is associated with an increased likelihood of smoking, but is not a consequence of it, and it is associated with lung cancer independent of its association with smoking. It is, therefore, possible that the observed association between smoking and lung cancer might simply reflect an association between aging and lung cancer.

Investigator error due to faulty data entry or inappropriate analysis is always possible, although it is unlikely in this study because the authors were well known for careful methods and proper analysis.

○ **21. Complete Table 3 by calculating the lung cancer mortality rates, rate ratios, and rate**

differences for each smoking category. What is the overall trend concerning lung cancer mortality rates? How would you interpret your findings for the rate ratio and rate difference categories?

Answer:

Daily number of cigarettes smoked	Mortality rate per 1000 person years	Rate Ratio	Rate Difference per 1000 person-years
0	0.07	Referent	Referent
1–14	0.57	8.1	0.50
15–24	1.39	19.8	1.32
25+	2.27	32.4	2.20
All smokers	1.30	18.6	1.23
Total	0.94		

Overall trend: Lung cancer mortality rates increase with an increase in the amount of cigarettes smoked per day (dose-response relationship).

Rate ratio: Comparing the categories 1–14 and 0 cigarettes a day yields a **rate ratio of 8.1** (0.57/0.07 = 8.1). Thus, the rate of lung cancer among people who smoke 1–14 cigarettes daily is **8.1 times higher** than the rate among nonsmokers. Put another way, people who smoke 1-14 cigarettes a day are **8.1 time more likely** than nonsmokers to develop lung cancer.

Rate difference: The excess death attributable to smoking increases from 0.50 to 2.20 per 1,000 person-years as the quantity of cigarettes smoked per day increases. Overall, smoking causes 1.23 excess deaths per 1000 people per year (1.30 ! 0.07).

> **Teacher's Note**:
> A rate is an expression of the frequency with which an event occurs in a defined population. Thus, rates can be calculated in cohort studies, which have well-defined populations, and not in case-control studies, which do not.

○ **22. Calculate the attributable risk percent for this study. What does the answer tell us?**

Answer:

Using the first formula: AR% = (1.30 − 0.07) ÷ 1.30 = 0.946 x 100% = 94.6%.

Thus, assuming the data are valid, about 95% of the deaths due to lung cancer in the exposure group (all smokers) may be attributable to smoking. This value can also be interpreted as the proportion of lung cancer deaths that could have been prevented among smokers if they had not smoked.

○ 23. Given your answer to question 21, how many deaths from lung cancer could have been avoided if no one had smoked?

Answer:

Overall, 95% of lung cancer deaths among smokers are attributable to smoking. Therefore, if no one had smoked, 95% of the 133 deaths (approximately 126) due to lung cancer would have been avoided.

○ 24. Given the data in Table 4, with which cause of death is smoking more strongly associated? Why?

Answer:

The rate ratio is the primary measure of association, and it indicates a much stronger association between smoking and lung cancer mortality than between smoking and cardiovascular mortality (18.6 vs. 1.3). These values suggest that smokers are 14 times more likely to die from lung cancer than from cardiovascular disease.

○ 25. From Table 4, calculate the population attributable risk percent for lung cancer mortality and cardiovascular disease mortality. What do these values tell us?

Answer:

PAR% for lung cancer = (0.94 − 0.07) ÷ 0.94 = 0.925 x 100% = 92.5%

PAR% for cardiovascular disease = (8.87 − 7.32) ÷ 8.87 = 0.174 x 100% = 17.4%

Therefore, 92.5% of all deaths due to lung cancer and 17.4% of all deaths due to cardiovascular disease in the study population are attributable to smoking. From a prevention perspective, the PAR% for a given exposure can be interpreted as the proportion of lung cancer cases in the entire study population that would have been prevented if exposure had not occurred.

Teacher's Note:
Whereas the **attributable risk percent** (see question 22) is the percentage of deaths among the *exposure group* that is attributable a certain risk factor, the **population attributable risk percent** is the percentage of deaths in the *entire study population* that is attributable to a certain risk factor. In this study, 95% of deaths from lung cancer and 23% of deaths from cardiovascular disease among the exposure group can be attributed to smoking. Note that the attributable risk percent and population attributable risk percent are similar because the prevalence of smoking in the study population is so high. If the smoking prevalence were lower, the PAR% would be lower. AR% is unaffected by prevalence.

○ **26. In the entire study population, how many deaths due to lung cancer per 1000 people per year (1000 person-years) can be attributed to smoking? How many deaths from cardiovascular disease?**

Answer:
To determine the number of deaths attributable to smoking, multiply the PAR% times the mortality rate per 1000 person-years in the study population for each disease:

Lung Cancer: 0.925 x (0.94) = 0.87 deaths per 1,000 person years
CV disease: 0.174 x (8.87) = 1.54 deaths per 1,000 person years

The number of smoking-related deaths per 1,000 people per year is greater for cardiovascular disease than for lung cancer, even though the relative risk (rate ratio) is considerably lower. Thus, if no one smoked, more deaths would be prevented from cardiovascular disease than from lung cancer.

○ **27. What do the data in Table 5 suggest regarding smokers, nonsmokers, and ex-smokers? What does this imply from a public health perspective?**

Answer:
Smokers have the highest risk of dying from lung cancer, and nonsmokers have the lowest. The risk for ex-smokers decreases with time after quitting, but even after 20 years, the risk is still nearly three times that of nonsmokers. From a public health perspective, smoking cessation efforts are worthwhile because risk is reduced, but smoking prevention efforts would be the most valuable, given the very low risk associated with not smoking at all.

○ 28. What is similar about the results of Doll's case-control study and Hill's cohort study? What is different? How would you account for this difference?

Answer:

Results from the studies are very consistent, exhibit a strong association between cigarette smoking and lung cancer, and include evidence of a dose-response relationship.

Compared with the rate ratios from the cohort study, the odds ratios in the case-control study consistently underestimate the strength of association. This discrepancy is most likely due to the fact that the controls in this study were also very likely to be smokers

○ 29. Compare case-control studies and cohort studies in terms of the following requirements:

Answer:

	Case-Control	Cohort
Sample size	Small	Large
Cost	Less	More
Study time	Short	Long

○ 30. Which type of study would be an advantage and which would be a disadvantage in investigations involving the following issues:

Answer:

	Case-Control	Cohort
Rare disease	Advantage	Disadvantage
Rare exposure	Disadvantage	Advantage
Multiple exposures	Advantage	Disadvantage

Multiple outcomes	Disadvantage	Advantage
Natural history of disease	Disadvantage	Advantage
Disease rate	Cannot measure	Advantage

○ 31. For which of the two types of studies are the following flaws likely to be more of a problem?

Answer:

	Case-Control	Cohort
Recall bias	Potential problem	Less of a problem
Loss to follow up	Advantage	Potential problem
Selection bias	Potential problem	Less of a problem

○ 32. Which type of study (case-control or cohort) would you have conducted first and why? Why conduct a second study? Which type of study should this be?

Answer:

A case-control study is quicker, easier, and less costly. If the case-control study results reveal a significant relationship between exposure and disease, then it is appropriate to do a second study to confirm the findings. Although a cohort study is more difficult and expensive to mount, as well as slower to yield results, it can confirm the results of previous studies, allow for the calculation of disease rates, and enable investigators to assess the natural progression from exposure to disease.

○ 33. Which of the following criteria for causality are met by the data from these two studies?

Teacher's Note
In general, these five criteria must be met to establish a cause-and-effect relationship:

- **Strength of association:** the relationship must be clear.
- **Consistency:** observation of the association must be repeatable in different populations at different times.
- **Temporality:** the cause (exposure) must precede the effect (disease).
- **Biologic gradient:** there must be a dose-response relationship.
- **Biologic plausibility:** the explanation must make sense biologically.

Answer:

	YES	**NO**
Strength of association	X	
Consistency with other studies	X	
Exposure precedes disease	X	
Dose-response effect	X	
Biologic plausibility		X