Lab exercise on Analytic Study Designs

Instructors' Guide

Goals: understand, differentiate, and identify relative strengths and weaknesses of the various types of epidemiologic study designs.

1. Read the following passages. Identify the study design and determine what measures of association can be estimated (justify your choices).

   a. Questionnaires were mailed to every 10th person listed in the city telephone directory. Each person was asked to list age, sex, smoking habits, and respiratory symptoms during the preceding seven days. About 20% of the questionnaires were completed and returned. About 10% of respondents reported having upper respiratory symptoms.

      Cross-sectional or prevalence study; can estimate prevalence and prevalence odds ratio.

   b. 1,500 employees of a major aircraft company were initially examined in 1951 and were classified by diagnostic criteria for coronary artery disease (CAD). New cases of CAD have been identified by examinations every three years and through death certificates. Attack rates in different subgroups have been computed.

      Prospective cohort study; can estimate incidence proportion (CI), incidence rate (ID), and their ratios (CIR, IDR) and differences (CID,IDD), the latter serving as measures of impact.

   c. A random sample of middle-aged sedentary adults were selected from four census tracts, and each person was examined for coronary artery disease (CAD). All persons without disease were randomly assigned to either a two-year program of aerobic exercise or a two-year arthritis-prevention non-aerobic exercise program. Both groups were observed semi-annually for incidence of CAD.

      Randomized controlled trial/intervention trial; can estimate incidence (CI and ID), relative risk (CIR,IDR), and measures of impact.

   d. A 39-year old woman presents with a mild sore throat, fever, malaise and headache and is treated with penicillin, for presumed streptococcal infection. She returns in a week with hypertension, fever, rash and abdominal pain. She responds favorably to chloramphenicol, after a diagnosis of Rocky Mountain spotted fever is made.

      This is a case report. It is not usually regarded as an "epidemiologic" study, since there is neither "population" nor comparison group. Presentation of an interesting

case can serve to alert other health care professionals of the possibility of misdiagnosing a potentially fatal disease.

e. 50 patients with thyroid cancer are identified and surveyed by patient interviews to identify previous radiation exposure.

This study is a case series. Since there is no comparison group of people without thyroid cancer, associations between thyroid cancer and various exposures cannot be examined except on the basis of information from outside of the study. Exposure prevalences among cases can be estimated and comparisons between different subtypes of the disease can be made. Case fatality can also be estimated by follow-up.

f. Patients admitted for carcinoma of the stomach and patients without a diagnosis of cancer are interviewed about their chewing tobacco history to assess the possible association of chewing tobacco and gastric cancer.

Case-control study. This study includes both diseased and nondiseased populations for comparisons of previous exposure to chewing tobacco. Incidence rates (either overall or by exposure status) of cancer cannot be determined without additional information.

g. Data on median income for households in census tracts within a large metropolitan county in the U.S. were obtained from the Census Bureau's Current Population Survey. Air pollution levels were measured in these same census tracts during a period of one-month. The data were analyzed using a geographic information system (GIS) to produce maps showing pollution and income levels by census tract.

Ecologic/community/group level/correlational/aggregate study; the census tract is the unit of analysis. The proportion of counties with high pollution levels can be estimated as a prevalence-type measure, but more often these data would be analyzed with cross-tabulation, correlation, regression, and graphical methods.

2. You have developed the hypothesis that automobile drivers who regularly sleep less than 6 hours/night have a higher incidence of fatal automobile accidents. Think about the design of a case-control study to test this hypothesis. Consider:

   a. How will you define a "case"? Are there any special considerations?

   b. Where would you find the cases?

   c. Name a suitable population from which to choose the controls.

   d. What major characteristic must you strive to measure similarly in each study participant?

e. What difficulties will be encountered in measuring this characteristic?

a. Cases can be drivers who die in an automobile crash. Criteria will need to be set to identify people who died as a result of the crash, even if death occurs only after a number of hours or days. Also, people who die from other causes (e.g., heart attack) while driving, which will often result in a crash, should presumably not be classified as cases. The purpose of restricting the study to fatal crashes is presumably to ensure complete reporting (many crashes without personal injury may go unreported) and perhaps also to set a threshold for severity. But even so there may be problems with heterogeneity, since the automobile and terrain may greatly influence whether a given crash is fatal. Questions: must the drivers be residents of NC? Must they have an NC driver's license?

b. Some states (including NC) have a fatal accident reporting system, which provides a great deal of useful information. Medical examiner records are another possible source. If cases are identified through death certificates, police records will be needed to differentiate between drivers and passengers.

c. From the state's motor vehicle licensing records, choose a sample of drivers in the state who have not had fatal accidents. The controls should be drivers (or have been drivers during the period when the deaths occurred), since that is the population from which the cases arose (the study base). There is the question of what should be done about cases who were driving despite having their license revoked. Should different controls be chosen for them?

d. Information about sleep, of course, as well as safety features of the automobile, passenger restraint systems in use, road conditions, weather, driver age, sex, and medical history, and any other factor that influences crash rates and probability of fatality.

e. Information on sleep for deceased drivers will certainly be difficult to obtain with any degree of accuracy. Since such information will need to come from proxy respondents (e.g., spouse, other family member, co-worker), should similar sources be used to obtain data for controls?

3. A case-control study is initiated to look at the association between alcohol consumption and breast cancer. The cases are 250 women diagnosed with breast cancer.

   a. What are the most important requirements for the control group?

   The controls must 1) be free of breast cancer and 2) represent the population from which the cases arose.

   b. What measure(s) can you estimate in order to quantify the strength of the association between breast cancer and alcohol consumption using the 250 cases and 250 controls?

Yes,The odds ratio (OR) can be estimated from the data. Since breast cancer is a rare disease, the OR will estimate the risk ratio and the rate ratio. (If cases only incident (newly occurring) cases were enrolled and controls were obtained from the source population by the "density method" then the OR would directly estimate the rate ratio even without the rare disease approximation.) An incidence difference (based on either incidence proportion or incidence rate) can be estimated only with additional information to estimate incidence.

4. What is the importance of randomization in an intervention trial, and what does it accomplish?

   Randomization is important because it increases the likelihood that differences in outcome between groups can be attributed to the treatments applied. If randomization is not used, the differences in outcome may be due to differing characteristics in the groups being compared. Randomization of a large number of persons achieves equal baseline distributions of known (e.g. age, sex, severity of disease), unknown, and unmeasurable risk factors. Of course, equivalence at baseline does not ensure equivalence throughout the trial.

5. What is meant by the phrase "ecologic fallacy"?

   The "ecologic fallacy" is the inference that the individuals in a group share the characteristics of a larger population. Groups of individuals may differ greatly from the larger group. An association between two group-level measures does not imply an association between the two measures at the individual-level. If, however, the factors under consideration belong to the group, rather than to the individuals so that inference is not being made to the individual-level, then the potential for an ecologic fallacy does not arise.

6. Different study designs have particular advantages and disadvantages. Contrast the case-control and cohort designs with respect to the following factors, for a study collecting new data.

   a. Cost.

   b. Time required for completion of study.

   c. Efficiency (in terms of information per subject)

   d. Design issues

   e. Difficulty in obtaining information.

   f. Bias

   g. What can be estimated

Case-Control	Cohort
a. Relatively inexpensive	**a. Relatively expensive**
b. Can draw cases from a larger population base and thus accrue incident cases more rapidly.	**b. Long follow-up period often needed**
c. More precision from fewer subjects	**c. Many subjects contribute little information**
d. Selection of control group determines the results	**d. May be difficult to recruit and retain subjects**
e. Often difficult to obtain accurate measures of previous exposures via recall or medical records.	**e. Losses due to attrition**
f. Vulnerable to bias in case ascertainment, selection of controls, exposure measurement.	**f. Vulnerable to bias in case ascertainment, differential detection of outcome, and attrition.**
g. Odds ratios (which may estimate IDR or CIR) for multiple exposures.	**g. Incidence, association, and impact for multiple disease outcomes.**