ERIC Notebook

Second Edition

Assessment of Diagnostic and Screening Tests

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Diagnostic and screening tests are used to detect the presence or severity of disease in individuals. Clinicians rely on these tests to make decisions in treating patients. Thus, it is important to assess the performance of diagnostic and screening tests before they are adopted in a clinical setting. The performance of any new diagnostic or screening test is assessed by comparing actual test results to the patient's true disease status (as assessed by a *gold standard*). The four measures used to evaluate a new test are the *sensitivity*, *specificity,* and *positive and negative predictive values.*

The gold standard

Gold standard is a term for the most definitive diagnostic procedure, e.g. microscopic examination of a tissue specimen, or the best available laboratory test, e.g. serum antibodies to HIV. Sometimes it can refer to a comprehensive clinical evaluation, e.g. clinical assessment of arthritis. Gold standard procedures can often be costly, invasive, and/or uncomfortable. New tests that are less invasive and less expensive are compared against gold standards to assess the new test's accuracy. A test which detects a marker in the blood for prostate cancer may not be as sensitive as taking a biopsy from the prostate itself, but the discomfort of biopsy may make the blood assay a better alternative.

Calculating the test results

A table like the one below is used to group individuals into one of four disease-test categories.

	Truth (Gold Standard)	
Results of test	Disease Present	Disease Absent
Positive Test	a = True Positive	b = False Positive
Negative Test	c = False Negative	d = True Negative

Prevalence of the disease =

$$\frac{\text{True positives}}{\text{Total population}} = \frac{(a+c)}{a+b+c+d)}$$

Sensitivity vs. specificity

Sensitivity and specificity are measures that assess the validity of diagnostic and screening tests. These measures reflect how well the test is

detecting the disease and classifying individuals into disease and non-disease groups.

Sensitivity (Se or Sn) describes how well the test detects disease in all who truly have disease, or the percent of diseased individuals who have positive test results

Specificity (Sp), describes how well the test is detecting non-diseased individuals as truly not having the disease, or the percent of non-diseased individuals who have

$$Se = \frac{True-positives}{True-positivies+False-negatives} \times 100$$

$$Se = \frac{a}{(a+c)} \times 100$$

$$Sp = \frac{True-negatives}{True-negativies+False-positives} \times 100$$

$$Sp = \frac{d}{(b+d)} \times 100$$

A highly sensitive test means that a large percent of people who have disease are classified correctly as having the disease. A highly specific test means that a large percent of individuals without disease are classified correctly as not having disease. An ideal test would be both highly sensitive and highly specific, where disease would be detected in 100% of those who truly have disease (100% sensitivity), and disease would be ruled out in 100% of those who are truly disease-free (100% specificity).

Example

If a test is 95% sensitive and 98% specific, then 5% of the diseased individuals will have negative test results (the test is incorrectly classifying 5% of the diseased individuals), and 2% of the disease-free individuals tested will have positive test results (the test is incorrectly classifying 2% of the disease-free individuals).

False-positives and false-negatives

A *false positive* is an individual who is incorrectly diagnosed as a case when, in fact, they do not have the disease. A *false negative* is an individual who is incorrectly diagnosed as a non-case, when in fact the person does have the disease.

100% - % sensitivity = % false negatives

100% - % specificity = % false positives

Positive and negative predictive values

The *positive predictive value* (PPV) is the percent of positive tests that are truly positive. The *negative predictive value* (NPV) is the percent of negative tests that are truly negative.

Like sensitivity and specificity, PPV and NPV also show how well the test is classifying individuals into disease and non-disease groups, but the denominator for PPV is the total number of persons who test positive (a+b), while that for NPV is the total number who test negative (c+d). A test with a high PPV value means that there is only a small percent of false-positives within all the individuals with positive test results. A test with a high NPV value means that there is only a small percent of false-negatives within all the individuals with negative test results.

$$PPV = \frac{True\ Positives}{Positive\ Tests} \times 100$$

$$PPV = \frac{a}{(a+b)} \times 100$$

$$NPV = \frac{True\ Negatives}{Negative\ Tests} x100$$

$$NPV = \frac{d}{(c+d)} \times 100$$

Example

A certain test, e.g. a stress ECG, which has a PPV of 90% and a NPV of 95% is used to screen 5,000 people for coronary heart disease. Forty percent of the individuals (2,000 people) have positive test results and 60% (3,000 people) have negative test results. But the gold standard for CHD found that 1,800 of those who tested positive (90% of 2,000) truly have CHD, and 2,850 of those who tested negative (95% of 3,000) are truly non-cases.

Pros and cons of specificity and sensitivity

Ideally, an investigator would prefer a diagnostic test that is both 100% sensitive and 100% specific. However, this scenario rarely occurs. It is important in clinical decision-making to know the sensitivity and specificity of the test you are conducting and to weigh the pros and cons of using tests with different levels of sensitivity and specificity. For instance, if a disease is not life threatening if left untreated, the costs of treatment are high, and invasive surgery is required, then a very specific diagnostic test is preferred over a more sensitive test. If the disease under study is life threatening if left untreated, and the survival rate is improved with immediate treatment, then the sensitivity of a diagnostic test is of greater importance than its specificity.

Example

Schizophrenia has a low prevalence in the U.S. at around 1%. A new diagnostic test that is 99% sensitive and 99% specific is used to screen 10,000 patients for schizophrenia. Of those 10,000, we would expect 100 to truly be suffering from schizophrenia, or 1% of our population. Of those 100, 99 (99% of 100) would have positive test results. Of the 9,900 who are truly without disease, 9801 (99% of 9,900) would be classified as disease-free. However, there would be 99 (9,900-9801) false positives. This test would give 198 (99+99) positive test results. Therefore, even with a test that is 99% sensitive and 99% specific, the PPV would only be 50% (99/198).

The prevalence affects the predictive values

The prevalence of a disease affects the PPV and NPV values. If a disease has a low prevalence and the test being used to assess disease in individuals is not 100% sensitive or 100% specific, as will most likely be the case, then false-positives may overwhelm the positive test results.

Terminology

False positive: An individual who is incorrectly diagnosed as having disease

False negative: An individual who is incorrectly diagnosed as not having disease

Gold standard: The most definitive diagnostic procedure to detect disease

Negative predictive value: The percent of negative tests that are truly negative

Positive predictive value: The percent of positive tests that are truly positive

Sensitivity: The percent of diseased individuals who have positive test results

Specificity: The percent of non-diseased individuals who have negative test results

Practice Questions.

A test is used to screen people for hepatitis B. The sensitivity of the test is 95% and the specificity of the test is 90%. Assume that the total number of persons being tested for hepatitis B is 50,000. Assume that the true prevalence of hepatitis B in the population is 100 per 50,000.

	Disease present	Disease absent
Positive test	a= true positive	b= false positive
Negative test	c= false negative	d= true negative

1) Calculate the number of true positives

2) Calculate the number of false positives

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

	Disease present	Disease absent
Positive test	a= true positive	b= false positive
Negative test	c= false negative	d= true negative

a+ b+ c+ d= 50,000
a+c= 100 (100 people have Hepatitis B)
b+d = Total – (a+c) = 50,000-100=49900

1) Calculate the number of true positives
Sensitivity = 95% = a / (a+c) a= (0.95)(a+c)= 0.95*100= 95 true positives

A sensitivity of 95% means that the test detects as positive 95% of the diseased individuals.

2) Calculate the number of false positives
Specificity = 90%= d / (d+b)
d=(0.90)(d+b) = 0.90*49900)=44910
b= (b+d)-d= (49900-44910)=4990 false positives

A specificity of 90% means that the test detects as negative 90% of the non-diseased individuals.

ERIC Notebook

Second Edition

Calculating Person-Time

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

What is person-time?

Person-time is an estimate of the actual time-at-risk – in years, months, or days – that all participants contributed to a study. In certain studies people are followed for different lengths of time, as some will remain free of a health outcome or disease longer than others. A subject is eligible to contribute person-time to the study only so long as that person does not yet have the health outcome under study and, therefore, is still at risk of developing the health outcome of interest. By knowing the number of new cases of the health outcome and the person-time-at-risk contributed to the study, an investigator can calculate the rate of the health outcome or disease, or how quickly people are acquiring the health outcome or disease.

Calculating rates

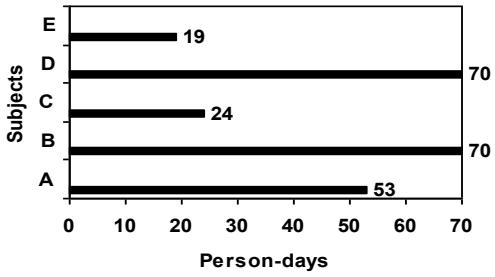
The rate is the number of new (incident) cases during study follow-up divided by the person-time-at-risk throughout the observation period.

$$\text{Rate} = \frac{\#\ of\ new\ cases}{total\ person\text{–}time\ at\ risk}$$

The denominator for a rate (person-time) is a more exact expression of the population at risk during the period of time when the change from non-disease to disease is being measured. The denominator for the rate changes as persons originally at risk develop the health outcome during the observation period and are removed from the denominator.

Calculating person-time for rates

Now suppose an investigator is conducting a study of the rate of second myocardial infarction (MI). He follows 5 subjects from baseline (first MI) for up to 10 weeks. The results are graphically displayed as follows:



The graph shows how many days each subject remained in the study as a non-case (no second MI) from baseline. From this graph the investigator can calculate person-

time. Person-time is the sum of total time contributed by all subjects. The unit for person-time in this study is person-days (p-d).

Time contributed by each subject:

Subject A: 53 days

Subject B: 70 days

Subject C: 24 days

Subject D: 70 days

Subject E: 19 days

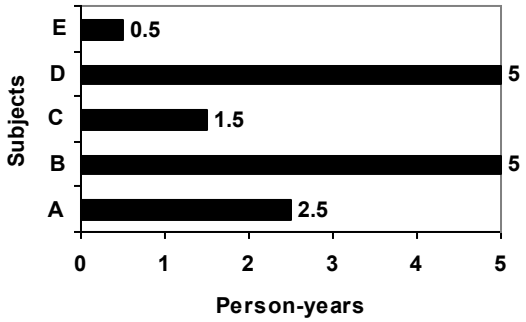
Total person-days in the study: 53+70+24+70+19=236 person-days

236 person-days (p-d) now becomes the denominator in the rate measure. The total number of subjects becoming cases (subjects A, C, and E) is the numerator in the rate measure. Therefore the rate of secondary MI is 3/(236 p-d), which is 0.0127 cases per person-day. By multiplying the numerator and denominator by 1000, the rate becomes 12.7 cases per 1000 person-days. The denominator, person-days, can be converted into other time units (such as hours or years) appropriate to the disease or health outcome being studied.

Secondary MI may be expressed in cases per person-year (p-y) by: (0.0127 cases/p-d) x (365 p-d/1 p-y) = 4.6 cases/p-y

Estimating when a person becomes a case

Now suppose an investigator is studying the rate of prostate cancer in men with a family history of prostate cancer. Subjects are examined once a year for up to five years. In order to calculate person-time when an investigator is only examining patients at specified intervals (once a year) the investigator must determine when a newly diagnosed case acquired the disease within the last year. In order to determine the amount of person-time adequately, an investigator may decide that the onset of prostate cancer occurred at the midpoint of the time interval between being disease free and becoming a case. This is because the investigator does not know precisely when subject A developed prostate cancer (just that it was sometime between exams two and three).



The following graph displays the amount of time until onset of prostate cancer for each subject.

Time contributed by each subject:

Subject A: 2.5 years

Subject B: 5 years

Subject C: 1.5 years

Subject D: 5 years

Subject E: 0.5 years

Total person-years in the study:

(2.5+5+1.5+5+0.5)=14.5 person-years

14.5 p-y is the denominator in the rate of prostate cancer. The rate is 3/(14.5 p-y), or 0.207 cases per p-y. By multiplying both the numerator and denominator by 1000 the rate becomes 207 cases per 1000 p-y.

Terminology

Rate: the number of new cases of disease during a period of time divided by the person-time-at-risk

Person-time: estimate of the actual time-at-risk in years, months, or days that all persons contributed to a study

Practice Questions

Answers are located at end of this notebook.

Researchers are studying the rate of developing asthma. The researchers enroll 100 participants who have been determined to not have asthma. The researchers plan to follow these participants over one year to see who develops asthma, beginning on January 1st. Participants visit a doctor monthly, at the end of the month, to determine if they have asthma. After one year, 5 of the participants have developed asthma. Two participants had asthma diagnosed at the end of March. Two participants had asthma diagnosed at the end of August. One participant had asthma diagnosed at the end of November.

1) How many person-months did the study participants contribute to the study, assuming that patients became cases of asthma on the last day of the month when they were diagnosed?

2) What is the rate of asthma cases in this study?

3) In this study, when were participants removed from the denominator of the rate?

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1)
(95 patients * 12 months)=1140
(2 patients * 3 months)=6
(2 patients*8 months)=16
(1 patient* 11 months)=11
Sum= 1140 + 6 + 16 + 11 =1173 person-months

2)
The one year rate = (# of new cases) / total person-time at risk = 5 cases / 1173 person-months = 0.0043

3)
Participants were removed when they were no longer at risk of the outcome, which was asthma. All participants began the study at-risk of developing asthma. Two patients were removed from the denominator of the rate at the end of March. Two participants were removed from the denominator of the rate at the end of August. One participant was removed from the denominator of the rate at the end of August. The remaining 95 asthma-free patients were removed from the denominator of the rate only at the very end of the study, which would have been December 31st.

ERIC Notebook

Second Edition

Case-Control Studies

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

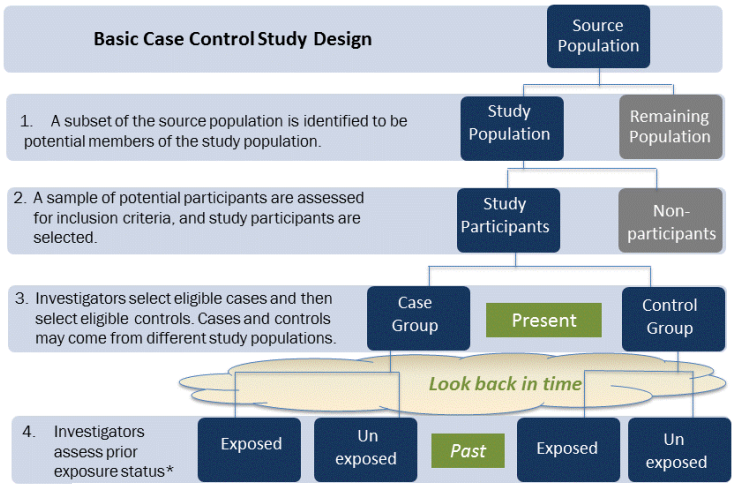
Case-control studies are used to determine if there is an association between an exposure and a specific health outcome. These studies proceed from effect (e.g. health outcome, condition, disease) to cause (exposure). Case-control studies assess whether exposure is disproportionately distributed between the cases and controls, which may indicate that the exposure is a risk factor for the health outcome under study. Case-control studies are frequently used for studying rare health outcomes or diseases.

Unlike cohort or cross-sectional studies, subjects in case-control studies are selected because they have the health outcome of interest (cases). Selection is not based on exposure status. Controls, persons who are free of the health outcome under study, are randomly selected from the population out of which the cases arose. The case-control study aims to achieve the same goals (comparison of exposed and unexposed) as a cohort study but does so more efficiently, by the use of sampling.

At baseline:

- Selection of cases and controls based on health outcome or disease status
- Exposure status is unknown

After cases and controls have been identified, the investigator determines the proportion of cases and the proportion of controls that have been



*Exposure at some specified point before disease onset

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

exposed to the exposure of interest. Thus, the denominators obtained in a case-control study do not represent the total number of exposed and non-exposed persons in the source population.

After the investigator determines the exposure, a table can be formed from the study data.

	Cases	Controls
Exposed	a	b
Unexposed	c	d

Measures of incidence in case-control studies

In case-control studies the proportion of cases in the entire population-at-risk is unknown, therefore one cannot measure incidence of the health outcome or disease. The controls are representative of the population-at-risk, but are only a sample of that population, therefore the denominator for a risk measure, the population- at-risk, is unknown. We decide on the number of diseased people (cases) and non-diseased people (controls) when we design our study, so the ratios of controls to cases is not biologically or substantively meaningful. However, we can obtain a valid estimate of the risk ratio or rate ratio by using the exposure odds ratio (OR).*

Odds of exposure among cases = a/c

Odds of exposure among controls = b/d

Diseased person-years

	Disease	No Disease
Exposed	a	n_1
Unexposed	c	n_2

RR = (a/n1)/(c/n2)

***Note:** Under some conditions, the odds ratio approximates a risk ratio or rate ratio. However, this is not always the case, and care should be taken to interpret odds ratios appropriately.*

Case-Control Study

	Cases	Controls
Exposed	a	b
Unexposed	c	d

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

If b and d (from the case-control study) are sampled from the source population, n1 + n2, then b will represent the n1 component of the cohort and d will represent the n2 component, and (a/n1)/(c/n2) will be estimated by (a/b)/(c/d).

Interpreting the odds ratio

The odds ratio is interpreted the same way as other ratio measures (risk ratio, rate ratio, etc.).

OR = 1	Odds of disease is the same for exposed and unexposed
OR > 1	Exposure increases odds of disease
OR < 1	Exposure reduces odds of disease

For example, investigators conducted a case-control study to determine if there is an association between colon cancer and a high fat diet. Cases were all confirmed colon cancer cases in North Carolina in 2010. Controls were a sample of North Carolina residents without colon cancer. The odds ratio was 4.0. This odds ratio tells us that individuals who consumed a high fat diet have four times the odds of colon cancer than do individuals who do not consume a high-fat diet. In another study of colon cancer and coffee consumption, the OR was 0.60. Thus, the odds of colon cancer among coffee drinkers is only 0.60 times the odds among individuals who do not consume coffee. This OR tells us that coffee consumption seems to be protective against colon cancer.

Types of case-control studies

Case-control studies can be categorized into different groups based on when the cases develop the health outcome and based on how controls are sampled. Some

case-control studies use prevalent cases while other case-control studies use incident cases. There are also different ways that cases can be identified, such as using population-based cases or hospital-based cases.

Types of cases used in case control studies

Prevalent cases are all persons who were existing cases of the health outcome or disease during the observation period. These studies yield a prevalence odds ratio, which will be influenced by the incidence rate and survival or migration out of the prevalence pool of cases, and thus does not estimate the rate ratio. Case control studies can also use incident cases, which are persons who newly develop the health outcome or disease during the observation period. Recall that prevalence is influenced by both incidence and duration. Researchers that study causes of disease typically prefer incident cases because they are usually interested in factors that lead up to the development of disease rather than factors that affect duration.

Selecting controls

Selection of controls is usually the most difficult part of conducting a case-control study. We will discuss 3 possible ways to select controls:

1. Base or case-base sampling
2. Cumulative density or survivor sampling
3. Incidence density or risk set sampling

Base sampling or case-base sampling

This sampling involves using controls selected from the source population such that every person has the same chance of being included as a control. This type of sampling only works with a previously defined cohort. In these case-control studies, the odds ratio provides a valid estimate of the risk ratio without assuming that the disease is rare in the source population.

Cumulative density sampling or survivor sampling

When controls are sampled from those people who remained free of the health outcome at the end of follow-up then we call the sampling cumulative density sampling or survivor sampling. Controls cannot ever have the outcome (become cases) when using this type of sampling. In these case-control studies, the odds ratio estimates the rate ratio only if the health outcome is rare, i.e. if the proportion of those with the health outcome among each exposure group is less than 10% (requires the rare disease assumption).

Incidence density sampling or risk set sampling

When cases are incident cases and when controls are selected from the at-risk source population at the same time as cases occur (controls must be eligible to become a case if the health outcome develops in the control at a later time during the period of observation) then we call this type of sampling incidence density sampling or risk set sampling. The control series provides an estimate of the proportion of the total person-time for exposed and unexposed cohorts in the source population. In these case -control studies, the odds ratio estimates the rate ratio of cohort studies, without assuming that the disease is rare in the source population.

Note that it is possible, albeit rare, that a control selected at a later time point could become a case during the remaining time that the study is running. This differs from case-control studies that use cumulative density sampling or survivor sampling, which select their controls after the conclusion of the study from among those individuals remaining at risk.

Selecting controls in a risk set sampling or incidence density sampling manner provides two advantages:

1. A direct estimate of the rate ratio is possible.
2. The estimates are not biased by differential loss to follow up among the exposed vs. unexposed controls.

For example, if a large number of smokers left the source population after a certain time point, they would not be available for selection at the end of the study – when controls would be selected in a study that uses cumulative density sampling or survivor sampling. This would give the

investigators biased information regarding the level of exposure among the controls over the course of the study.

Source populations for case-control studies

Source populations can be restricted to a population of particular interest, e.g. postmenopausal women at risk of breast cancer. This restriction makes it easier to control for extraneous confounders in the population. Controls should represent the restricted source population from which cases arise, not all non-cases in the total population. The cases in the study do not have to include all cases in the total population.

Sources of cases

- Cases diagnosed in a hospital or clinic
- Cases entered into a disease registry, e.g. cancer, birth defects, deaths
- Cases identified through mass screening, e.g. hypertensives, diabetics
- Cases identified through a prior cohort study, e.g. lung cancers in an occupational asbestos cohort

Sources of controls

- Population controls are non-cases sampled from the source population giving rise to cases. This is the most desirable method for selecting controls. Sampling randomly from census block groups, or a registry such as the Department of Motor Vehicles (of adults who are able to drive) are examples of ways to find and recruit population-based controls.
- Neighborhood or friend controls are appropriate for selection as controls if these individuals would be included as cases if they developed the health outcome of interest. It is not appropriate to select neighbors or friends as controls if they share the exposure of interest.
- Hospital controls - There are certain problems with hospital controls in that they may not be from the same source population from which the cases arose. Hospital controls may not be representative of the exposure prevalence in the source population of cases, e.g. there may be a higher prevalence of smokers in hospitals. Hospital controls also may have diseases resulting from the exposure of interest, e.g. the exposure (smoking) is related to the disease of interest (cancer) and to heart and lung diseases from which the controls may be suffering.
- Controls with another disease - However if the study is on lung cancer, for example, it is essential to exclude cancers known or suspected to be related to the study exposure of interest. These controls also share some of the same problems as hospital controls.

Advantages of case-control studies

Case-control studies are the most efficient design for rare diseases and require a much smaller study sample than cohort studies. Additionally, investigators can avoid the logistical challenges of following a large sample over time. Thus, case-control studies also allow more intensive evaluation of exposures of cases and controls. Case-control studies that use incidence density sampling or risk set sampling yield a valid estimate of the rate ratio derived from a cohort study if incident cases are studied and controls are sampled from the risk set of the source population. If properly performed (i.e. appropriate sampling), case-control studies provide information that mirrors what could be learned from a cohort study, usually at considerably less cost and time.

Disadvantages of case-control studies

Case-control studies do not yield an estimate of rate or risk, as the denominator of these measures is not defined. Case-control studies may be subject to recall bias if exposure is measured by interviews and if recall of exposure differs between cases and controls. However, investigators may be able to avoid this problem if historical records are available to assess exposure. Choosing an appropriate source population is also difficult and may contribute to selection bias. Case-control studies are not an efficient means for studying rare exposures (less than 10% of controls are exposed) because very large numbers of cases and controls are needed to detect the effects of rare exposures.

Terminology

Cohort studies: An observational study in which subjects are sampled based on the presence (exposed) or absence(unexposed) of a risk factor of interest. These subjects are followed over time for the development of a health outcome of interest.

Cross-sectional studies: An observational study in which subjects are sampled at one point in time, and then the associations between the concurrent risk factors and health outcomes are investigated.

Exposure odds ratio (OR): the odds of a particular exposure among persons with a specific health outcome divided by the corresponding odds of exposure among persons without the health outcome of interest. Yields a valid estimate of the incidence rate ratio or risk ratio derived from a cohort study, depending on control sampling.

Incident case: a person who is newly diagnosed as a case.

Prevalent case: a person who has a health outcome of interest that was diagnosed in the past.

Risk ratio (RR): the likelihood of a particular health outcome occurrence among persons exposed to a given risk factor divided by the corresponding likelihood among unexposed persons.

Source population: the population out of which the cases arose.

From: Medical Epidemiology, R.S. Greenberg, 1993, 1996.

Practice Questions

Answers are located at the end of this notebook

1) Researchers conduct a case-control study of breast cancer, using incident cases. The researchers find out that 90% of the cases had taken hormonal contraceptives in the past. Should the researchers conclude that hormonal contraceptives increase the risk of developing breast cancer?

2) Researchers conduct a case-control study of pancreatic cancer. The study included 200 cases and 200 controls. Of the cases, 80% reported they smoked cigarettes. Among the controls, 50% reported they smoked cigarettes.

a) Prepare a 2x2 table with these data

b) Calculate the exposure odds ratio

c) Interpret the exposure odds ratio in a sentence

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-.2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. No, The information provided in this question is only for cases. No information was given in the question about the control group that was studied. To assess if exposure to hormonal contraceptives was associated with the risk of breast cancer, information would be required on the proportion of control subjects who had taken hormonal contraceptives and the researchers would need to calculate the exposure odds ratio.

2.

a) 2x2 table

	Cases	Controls	
Exposed	160 (a)	100 (b)	260
Unexposed	40 (c)	100 (d)	140
	200	200	400

b) Exposure odds ratio = (a/c) / (b/d) = (a*d)/(c*b) = (160*100) / (40*100) = 4.0

c) An odds ratio of 4.0 means that the odds of smokers being a case are 4 times the odds of non-smokers being a case.

ERIC Notebook

Second Edition

Causality

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

The primary goal of the epidemiologist is to identify those factors that have a causal impact on development or prevention of a health outcome, thereby providing a target for prevention and intervention. At first glance, causality may appear to be a relatively simple concept to define. However, adequately distinguishing causal agents from non-causal agents is not an easy task from an epidemiologic perspective. Unfortunately, there is no elementary parameter that can be measured to provide a definitive answer when determining causality. Rather, there are a series of criteria that have been developed and refined over the years that now serve as the guideline for causal inference. The most important point to remember is that causality is not determined by any one factor, rather it is a conclusion built on the preponderance of the evidence.

Epidemiologist Austin Bradford Hill is credited with identifying the nine factors that constitute the current standard for determining causality (1965). In his article, Hill expanded upon criteria that had previously been set forth in the report *Smoking and Health* (1964) by the United States Surgeon General. Below is a discussion of the nine criteria defined by Hill to be utilized in the determination of causality.

It is important to note that satisfying these criteria may lend support for causality, but failing to meet some criteria does not necessarily provide evidence against causality, either. Hill's causal criteria should be viewed as guidelines, not as a "checklist" that must be satisfied for a causal relationship to exist.

Hill's causal criteria

Strength of association

Strength of association between the exposure of interest and the outcome is most commonly measured via risk ratios, rate ratios, or odds ratios. Hill believed that causal relationships were more likely to demonstrate strong associations than were non-causal agents. Smoking and lung cancer is a perfect example where risk ratios, rate ratios, and odds ratios are in the 20 to 40 range when comparing smokers to non-smokers. However, weak associations as demonstrated by the risk ratio, rate ratio, or odds ratio should not be taken as an indication of non-causality. This is particularly true when the outcome of interest is common.

An example of a common outcome that exhibits a weak association to smoking is cardiovascular disease (CVD). Yet even with a weak association, evidence supports the casual nature between smoking and the development of CVD. Furthermore, one should not assume that a strong association alone is indicative of causality, as the presence of strong confounding may erroneously lead to a strong causal association.

Consistency of data

This tenant refers to the reproducibility of results in various populations and situations.

Consistency is generally utilized to rule out other explanations for the development of a given outcome. It should also be noted that a lack of consistency does not negate a causal association as some causal agents are causal only in the presence of other co-factors. In general, the greater the consistency, the more likely a causal association.

Specificity

This criterion has been proven to be invalid in a number of instances, with smoking being the primary example. Evidence clearly demonstrates that smoking does not lead solely to lung carcinogenesis but to a myriad of other clinical disorders ranging from emphysema to heart disease. On the other hand, there are certain situations where a 1 to 1 relationship exists, such as with certain pathogens which are necessary to produce a specific disease. Tuberculosis is a good example.

Temporality

This criterion has been identified as being the most likely to be the sine qua non for causality, i.e. it is absolutely essential. For an agent to be causal, its presence must precede the development of the outcome. Lack of temporality rules out causality. An example found in the literature is the relationship between atrial fibrillation (AF) and pulmonary embolism. Current wisdom supports that pulmonary embolism causes atrial fibrillation, however more recent evidence and plausible biological hypothesis suggest that the reverse could be true. Determining the proper course of care may hinge upon discovering if pulmonary emboli can indeed precede and thus perhaps cause the development of atrial fibrillation.

Dose-response

The presence of a dose-response relationship between an exposure and outcome provides good evidence for a causal relationship; however, its absence should not be taken as evidence against such a relationship. Some diseases or health outcomes do not display a dose-response relationship with a causal exposure. They may demonstrate a threshold association where a given level of exposure is required for disease or health outcome initiation, and any additional exposure does not affect the outcome.

Biological plausibility

Support for this criterion is generally garnered in the basic science laboratory. It is not unusual for epidemiological conclusions to be reached in the absence of evidence from the laboratory, particularly in situations where the epidemiological results are the first evidence of a relationship between an exposure and an outcome. However, one can further support a causal relationship with the addition of a reasonable biological mode of action, even though basic science data may not yet be available.

Coherence

This term represents the idea that, for a causal association to be supported, any new data should not be in opposition to the current evidence, that is, providing evidence against causality. However, one should be cautious in making definite conclusions regarding causation, since it is possible that conflicting information is incorrect or highly biased.

Experimental evidence

Today's understanding of Hill's criteria of experimental evidence results from many areas: the laboratory, epidemiological studies, and preventive and clinical trials. Ideally, epidemiologists would like experimental evidence obtained from a well-controlled study, specifically randomized trials. These types of studies can support causality by demonstrating that "altering the cause alters the effect".

Analogy

This is perhaps one of the weaker of the criteria in that analogy is speculative in nature and is dependent upon the subjective opinion of the researcher. An example of an analogy is that while infection may cause a fever, not all fevers are due to infection. Absence of analogies should not be taken as evidence against causation.

Other considerations

In addition to assessing the components of Hill's list, it is also critically important to have a thorough understanding of the literature to determine if any other plausible explanations have been considered and tested previously.

Additional models for causality

In additional to Hill's guidelines for causality, several recent models for understanding causality have been developed. These include Kenneth Rothman's component cause theory, counterfactual models, and directed acyclic graphs. We provide brief descriptions below but refer you to the suggested readings section for more information.

Component causes

Kenneth Rothman described the circumstances leading to a health outcome as being parts of one pie chart, or a "causal pie." Without each component in place, the disease or health outcome would not have occurred at that specific point in time.

Example

If you were to slip on an icy sidewalk and break your wrist, there may be a number of factors that contributed to that outcome. First, the fall directly leads to a broken wrist, and this was a necessary component of the outcome. However, which factors led to you walking on the ice and having a fall in the first place, and what other factors influenced your broken wrist? Maybe you were wearing poor footwear and were tugged by your dog who was chasing a squirrel. Maybe you didn't receive enough calcium in your diet, developed osteoporosis, and the fall would not have broken your wrist without weakened bones. Each of these components ultimately led to the outcome.

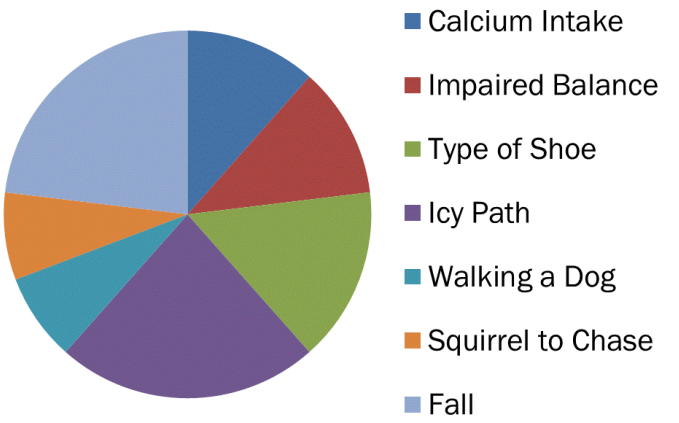


Figure: Causal Pie

Rothman's component causes theory is one way to consider all factors involved in the development of a health outcome.

Counterfactual models

Many statistical models that are used to adjust for confounding are based on counterfactual thinking. These models are based on comparing an exposed group of people to a fictional group of people who are exactly the same except they are unexposed to the key variable. These models try to answer the question: "If this one experience or exposure in the past did not happen to an individual, how would it impact that person's health outcome today?" Of course, this is an impossible situation; we cannot go back in time and change an individual's exposure status and track both outcomes over time. However, using two very similar populations of people – one group that is exposed and another that is unexposed – it is possible to estimate that very counterfactual situation on a group level. Many statistical modeling programs that adjust for potential confounders are modeling a counterfactual scenario to produce a less biased measure of association.

Directed acyclic graphs (DAGs)

DAGs are one method used to create a conceptual diagram that maps the relationships between the main exposure, outcome of interest, and all potential confounders for a given study. Through a DAG analysis, specific rules are followed to determine if confounding might be present for a given research question. Once confounders have been identified and adjusted for, a less biased measure of association can be obtained.

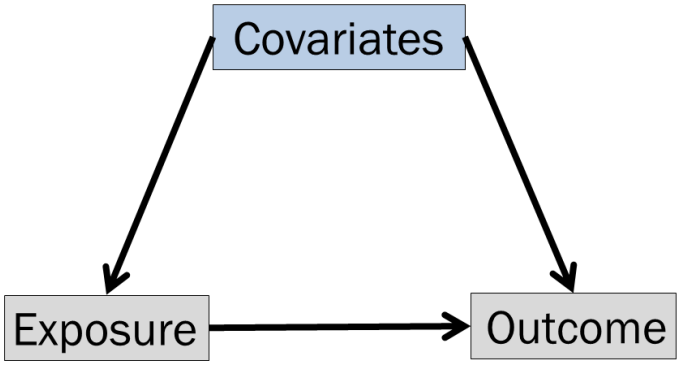


Figure: Basic Directed acyclic graph (DAG)

Practice Questions

Answers are located at end of this notebook.

1) Researchers conducted a cohort study of the association between air pollution and asthma. The rate ratio was 8.0, when comparing those exposed to high levels of air pollution with those exposed to low levels of air pollution. Which of the following issues should the researchers consider when making their study conclusions and when thinking about causality? Choose all that apply.

a) The rate ratio of 8.0 indicates a strong association, which lends support for causality
b) Strong confounding may actually be causing the strong association seen
c) Other studies of the same exposure—health outcome association reported rate ratios in the range of 1.5- 3.0, less than the rate ratio of 8.0 seen in this study
d) The temporal sequence of the exposure and outcome should be known in order to draw accurate conclusions

2) Considering the component causes model of causality developed by Kenneth Rothman, which of the following factors may play a role in the cause of a car accident? Choose all that apply.

a) The road conditions
b) The condition of the car
c) The amount of sleep that the driver of the car got the previous night
d) The level of distraction of the driver
e) The driver's overall health

References and Suggested Readings

Flegel KM. When atrial fibrillation occurs with pulmonary embolism, is it the chicken or the egg? *CMAJ*. 160 (8):1181-2, 1999.

Greenland S, Holland PW, Mantel N, Wickramaratne PJ and Holford TR. Confounding in Epidemiologic Studies. *Biometrics*. 45(4):1309-1322

Hill AB. The environment and disease: association or causation? *Proc R Soc Med.* 58:295-300, 1965.

Mengersen KL. Merrilees MJ. Tweedie RL. Environmental tobacco smoke and ischaemic heart disease: a case study in applying causal criteria. *Int Arch of Occup & Env Health*. 72 Suppl:R1-40, 1999.

Ridgway D. The logic of causation and the risk of paralytic poliomyelitis for an American child. *Epidemiology & Infection*. 124(1):113-20, 2000.

Rothman KJ. Causes. *Am J Epidemiol* 104:587–92, 1976.

Rothman KJ, Greenland S, Poole C, and Lash TL. Causation and causal inference. In: *Modern Epidemiology (3e)*. Edited by: Rothman KJ, Greenland S, and Lash TL. Philadelphia: Lippencott-Raven Publishers; 2008:5-31.

Shrier I and Platt RW. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology* 2008, 8:70.

United States Department of Health, Education, and Welfare. *Smoking and health:* Report of the Advisory Committee to the Surgeon General of the Public Health Service. Washington, D.C Government Printing Office, 1964 PHS Publ. No. 1103.

Weed DL. On the use of causal criteria. *Int J of Epidemiology*. 26(6):1137-41, 1997.

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Sandra Demming, MPH, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1) All answer choices are correct. When making their study conclusions and considering causality, the researchers would need to consider all of the issues listed here as well as additional issues. It is true that a rate ratio of 8.0 indicates a strong association. Such a large rate ratio would lend support for causality if this rate ratio is indeed correct. However, if strong confounding is present in a study then a strong association may actually be seen in error so the researchers would need to consider the presence of and extent of confounding. It would also be important for the researchers to consider previous studies of the same exposure—health outcome association. If all previous literature had reported rate ratios in the range of only 1.5- 3.0, then the researchers would want to be especially careful in determining if their results were really accurate and non-biased. Finally, the timing of the exposure would also be important to consider. Temporality is a key criterion because, for a relationship to be causal, the exposure must precede the outcome. The researchers would have needed to carefully set up their study to ensure that they were studying participants who were first exposed to air pollution and then went on to develop asthma, rather than studying participants who already had asthma prior to being exposed to high levels of air pollution.

2) All answer are correct. Kenneth Rothman's component cause theory is another way to understand causality. It is possible that a particular car accident may have occurred only because all of these factors occurred. Perhaps if even one of these factors had not occurred, the car accident would not have occurred. For example, this

fictitious car accident may have occurred on a day with wet road conditions. The car's tires may have not been in good condition, leading to a further decrease in stopping time on wet roads. The driver may have stayed up late the previous night, leading to a slower reaction time due to not getting sufficient sleep. The driver's reaction time might still have been quick enough to avoid the accident, however, due to holding a cup of coffee (a distraction) and due to having arthritis (a health problem), the driver was unable to react in time to steer the car away from a collision. Furthermore, chance may also have played a role in the accident. It is possible that had the driver not been on a particular portion of the road, at a particular time, that the accident would not have happened. All of these factors may have contributed in unison to lead to the occurrence of the car accident.

ERIC Notebook

Second Edition

Cohort Studies

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

A cohort study is a type of epidemiological study in which a group of people with a common characteristic is followed over time to find how many reach a certain health outcome of interest (disease, condition, event, death, or a change in health status or behavior). A cohort is defined as a group of persons, usually 100 or more in size, who share a common characteristic, e.g. smokers, workers in a lead smelter, people born in the same year, or all enrollees of a specific health insurance plan Cohort studies compare an exposed group of individuals to an unexposed (or less exposed) group of individuals to determine if the outcome of interest is associated with exposure. There are two types of cohort studies: prospective and retrospective (or historical) cohorts. Prospective studies follow a cohort into the future for a health outcome, while retrospective studies trace the cohort back in time for exposure information after the outcome has occurred. Both types of cohort studies are also referred to as longitudinal or follow-up studies.

Establishing the cohort

The investigator controls the selection of the cohort. The investigator may choose a cohort based on age, location, exposure to a certain working environment, or some other common characteristic. Cohorts may be selected on the basis of exposures known at baseline, e.g. smokers vs. nonsmokers. Alternatively, cohorts may be divided into exposure categories once baseline measurements of a defined population are made. For example, the Framingham Cardiovascular Disease Study (CVD) used baseline measurements to divide the population into categories of CVD risk factors.

For instance, an investigator wants to study whether exposure to military aircraft engine noise is a risk factor for hearing loss. The cohort this investigator would want to establish should be composed of two groups of military personnel: one exposed to engine aircraft noise (the group under study) and the other unexposed to engine aircraft noise (a comparison group). The unexposed group should be representative of the exposed group on all factors except exposure.

The cohort at baseline

After the cohort of study subjects is established, their individual exposures of interest are identified at baseline (through interviews, questionnaires, bioassays, medical records, etc.). Subjects with the outcome of interest at baseline are excluded. Therefore, all members of the cohort are at risk of developing the outcome at the beginning of observation.

Following the last example, anyone in the cohort of military personnel with a specified hearing loss at baseline would be excluded from the cohort and would not be followed.

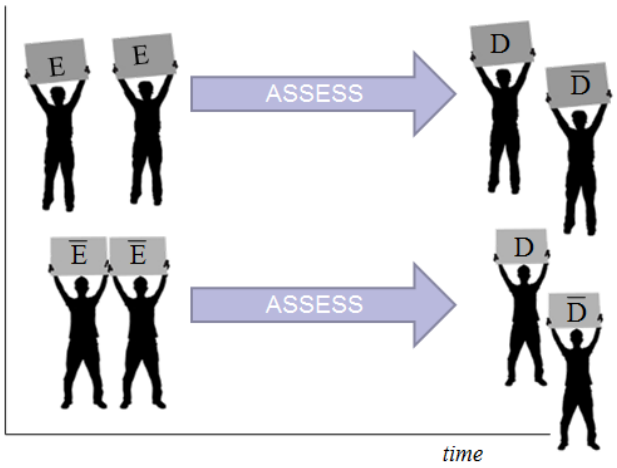
Following the cohort

The cohort is then followed over time for new occurrences of the outcome of

interest, in the above example, hearing loss. In a prospective, or concurrent, cohort study baseline exposure is assessed at the beginning of the study and the cohort is followed into the future. In a retrospective, or historical cohort study, baseline exposure is assessed at some point in the past through historical records, e.g. health records for a cohort of factory workers may provide exposure and outcome information up to the present.

Cohort	Baseline	Follow-up
Prospective	Assessed at beginning of study	Followed into the future for outcome
Retrospective	Assessed at some point in the past via historical records	Outcome has already occurred and is assessed via historical records

Cohorts are followed over time to the end of follow-up. Occurrence of the outcome of interest may be determined via interviews with members of the cohort and/or family members, or by viewing health and/or work records to conclude the study.



The basic design of a cohort study from beginning of the study to end of follow-up. E = exposed, Ē = not exposed, D = diseased and D̄ = not diseased.

Evaluation of the results

During the follow-up period the investigator counts the number of subjects who develop the outcome of interest. This count is the numerator for a calculation of risk, also referred to as **cumulative incidence** or **incidence proportion**. The number of persons at risk at baseline is the denominator.

Risk (also called Cumulative Incidence or Incidence Proportion) = new occurrences of the outcome / population-at-risk at baseline

Two risks can be compared to provide a risk ratio. The reference group is a comparable unexposed cohort. The index group is the exposed cohort. The risk ratio is computed by dividing the risk in the exposed group by the risk in the unexposed group. The risk ratio gives a relative measure of the increase or decrease in incidence between the exposed and unexposed groups.

$$\text{Risk Ratio} = Risk_{exposed}/Risk_{unexposed}$$

As with risk, **an incidence rate measure (IR)** is calculated with new occurrences of the outcome as the numerator. An incidence rate is also called incidence density. However, in an IR calculation the denominator is person-time (days, months, or years) at risk during follow-up. Person-time is measured by summing the total time each member of the cohort was free of the outcome of interest and thus contributed to person-time-at-risk during the follow-up period. The IR measures the rapidity of occurrence of new health outcomes in the population.

$$\text{Incidence rate} = \frac{new\ occurrences\ of\ the\ outcome}{person-time\ at\ risk}$$

Two IRs may also be compared to find the relative increase or decrease in the rate of health outcome occurrence between the exposed and unexposed groups. This relative measure is called the **incidence rate** ratio (IRR) or the rate ratio.

$$\text{Incidence rate ratio} = \frac{IR_{exposed}}{IR_{unexposed}}$$

Incidence measures between exposed and unexposed cohorts can also be subtracted from one another to find the difference between the two

measures. This measure is referred to as a **risk difference** or a **rate** difference.

$$\text{Risk Difference} = Risk_{exposed} - Risk_{unexposed}$$

$$\text{Rate Difference} = I_{exposed} - I_{unexposed}$$

Exposure may be a risk factor or a preventive factor in the development of the outcome of interest. When exposure is preventive, the risk ratio or rate ratio will be less than one.

Advantages of a cohort study

A cohort study can be used to directly measure the risk and rate of a health outcome occurrence over time. Cohort studies are an efficient means of studying rare exposures (e.g. gasoline fumes, as discussed in the next paragraph), in contrast to case-control studies, which tend to be better for rare outcomes. Cohort studies also allow the investigator to assess multiple outcomes of a single exposure.

> A cohort study would be the most efficient means of studying the effects of long-term exposure to gasoline fumes. The cohort would consist of individuals who are exposed daily to gasoline fumes (auto mechanics, gas station attendants, sea crewman on tankers, etc.). By studying this group
> of individuals, the investigator can better determine the direct effects of long-term, regular gasoline inhalation. Also, by conducting a cohort study, an investigator could determine if gasoline inhalation causes many different health outcomes (e.g., different types of cancer and respiratory illnesses).

Additional advantages of cohort studies

Cohort studies establish temporal relationships between exposure and outcome. Exposure clearly precedes the outcome because the population under study at baseline is free of the outcome of interest. Cohort studies also avoid recall bias (as the exposure is determined before the outcome, one's health outcome or disease state won't affect how accurately one recalls exposure levels), as well as, survival bias (duration of disease influencing exposure measurements). Therefore, cohort studies are the best observational study design used to help establish cause and effect relationships.

Disadvantages of cohort studies

Cohort studies often require large sample sizes, especially when the outcome is rare, defined as less than 1 event per 1000 person- years (e.g., all specific cancers). Therefore, cohort studies tend to be expensive and time-consuming. When there are *losses to follow-up* (individuals who leave the cohort before the end of follow-up) biases may occur. Thus, individuals who leave the cohort prematurely may have a different baseline risk than the members who remain in the cohort throughout the entire length of follow-up. Therefore, the study may not be *generalizable* to the original target population, but only to those who remained under investigation throughout the length of the study. Also, any differences in the quality of measurement of exposure or disease between exposed and non-exposed cohorts may introduce *information bias* and thereby distort the results.

Practice Questions

Answers are located at end of this notebook.

Researchers are conducting a prospective cohort study of the association between being an office worker who uses a computer daily and carpal tunnel syndrome (a hand/ arm nerve condition).

1) Researchers choose a group of office workers from one company to be their "exposed" group. Which of the following "unexposed or less exposed" comparison groups would be the <u>best</u> choice?

a) A group of high school students, who use computers daily

b) Another group of office workers at a different company, who do not use computers daily

c) Another group of office workers at the same company, who do not use computers daily

d) A group of professional golfers, who use their hands/ arms intensively for their sport

2) Who should be excluded from the cohort (not included in the study)?

3) A total of 300 exposed and 300 unexposed participants are enrolled and followed for 10 years. A total of 25 exposed and 17 unexposed had the outcome of interest over the follow-up period.

a) What is the risk of developing carpal tunnel syndrome among the exposed?

b) What is the risk of developing carpal tunnel syndrome among the unexposed?

c) What is the risk ratio for developing carpal tunnel syndrome?

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1) The best answer choice is answer c, another group of office workers at the same company, who do not use computers daily. The unexposed group should be representative of the exposed group on all factors except exposure. If the researchers used students as a comparison group, they would very likely be studying a completely different age-group of participants who would likely be quite different from a population of office workers. Similarly, if the researchers used the golfers as a comparison group, they would likely be studying a group of participants that vastly differed from office workers. Studying another group of office workers from a different company would be a fairly good choice since then "unexposed" office workers would be compared with the "exposed" office workers. However, the best choice would be to compare office workers at the same company, who live in the same area, looking at those who used a computer daily versus those who did not.

2) The researchers should exclude any participants who already have carpal tunnel syndrome at the start of the study.

3a) The 10-year risk is 25/300=0.083
3b) The 10-year risk is 17/300=0.057
3c) Risk exposed / Risk unexposed = 0.083 / 0.057 = 1.47 The risk of developing carpal tunnel syndrome among office workers exposed to daily computer use is 1.47 times the risk among office workers who do not use a computer daily.

ERIC Notebook

Second Edition

Common Measures and Statistics in Epidemiological Literature

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

For the non-epidemiologist or non-statistician, understanding the statistical nomenclature presented in journal articles can sometimes be challenging, particularly since multiple terms are often used interchangeably, and still others are presented without definition. This notebook will provide a basic introduction to the terminology commonly found in epidemiological literature.

Measures of frequency

Measures of frequency characterize the occurrence of health outcomes, disease, or death in a population. These measures are descriptive in nature and indicate how likely one is to develop a health outcome in a specified population. The three most common measures of health outcome or frequency are risk, rate, and prevalence.

Risk

Risk, also known as incidence, cumulative incidence, incidence proportion, or attack rate (although not really a rate at all) is a measure of the probability of an unaffected individual developing a specified health outcome over a given period of time. For a given period of time (i.e.: 1 month, 5 years, lifetime):

A 5-year risk of 0.10 indicates that

$$\text{Risk} = \frac{\#\ of\ new\ cases}{total\ \#\ individuals\ at\ risk}$$

an individual at risk has a 10% chance of developing the given health outcome over a 5-year period of time.

Risk is generally measured in prospective studies as the population at risk can be defined at the start of the study and followed for the development of the health outcome. However, risk cannot be measured directly in case-control studies as the total population at risk cannot be defined. Thus, in case-control studies, a group of individuals that have the health outcome and a group of individuals that do not have the health outcome are selected, and the odds of developing the health outcome are calculated as opposed to calculating risk.

$$\text{Odds} = \frac{\#\ individuals\ with\ the\ health\ outcome}{\#\ individuals\ without\ the\ health\ outcome}$$

Rate

A rate, also known as an incidence rate or incidence density, is a measure of how quickly the health outcome is occurring in a population. The numerator is the same as in risk, but the denominator includes a measure of person-time, typically person-years. (Person-time is defined as the sum of time that each at-risk individual contributes to the study).

$$\text{Rate} = \frac{\#\ of\ new\ cases}{total\ person\text{-}time\ at\ risk}$$

Thus a rate of 0.1 case/person-years indicates that, on average, for every 10 person-years (i.e.: 10 people each followed 1 year or 2 people followed

for 5 years, etc.) contributed, 1 new case of the health outcome will develop.

Prevalence

Prevalence is the proportion of a population who has the health outcome at a given period of time. Prevalence is generally the preferred measure when it is difficult to define onset of the health outcome or disease (such as asthma), or any disease of long duration (e.g. chronic conditions such as arthritis). A limitation of the prevalence measure is that it tends to favor the inclusion of chronic diseases over acute ones. Also, inferring causality is troublesome with prevalence data, as typically both the exposure and outcome are measured at the same time. Thus it may be difficult to determine if the suspected cause precedes the outcome of interest.

$$\text{Prevalence} = \frac{\#\ of\ affected\ individuals}{total\ \#\ of\ individuals\ in\ the\ population}$$

Thus a population with a heart disease prevalence of 0.25 indicates that 25% of the population is affected by heart disease at a specified moment in time.

A final note, risk and rates can also refer to deaths in a population and are termed mortality and mortality rate, respectively.

Measures of association

Measures of association are utilized to compare the association between a specific exposure and health outcome, They can also be used to compare two or more populations, typically those with differing exposure or health outcome status, to identify factors with possible etiological roles in health outcome onset. Note that evidence of an association does not imply that the relationship is causal; the association may be artifactual or non-causal as well. Common measures of association include the risk difference, risk ratio, rate ratio and odds ratio.

Risk difference

Risk difference is defined as

$$= Risk_{exposed} - Risk_{unexposed}$$

$$\frac{\#\ cases\ in\ exposed\ group}{total\ \#\ at\ risk\ in\ exposed\ group} - \frac{\#\ cases\ in\ control\ group}{total\ \#\ at\ risk\ in\ control\ group}$$

The risk difference, also know as the attributable risk, provides the difference in risk between two groups indicating how much excess risk is due to the exposure of interest. A positive risk difference indicates excess risk due to the exposure, while a negative result indicate that the exposure of interest has a protective effect against the outcome. (Vaccinations would be a good example of an exposure with a protective effect). This measure if often utilized to determine how much risk can be prevented by an effective intervention.

Risk ratio and rate ratio

Risk ratios or rate ratios are commonly found in cohort studies and are defined as: the ratio of the risk in the exposed group to the risk in the unexposed group or the ratio of the rate in the exposed group to the rate in the unexposed group

$$\text{Risk Ratio} = Risk_{exposed}/Risk_{unexposed}$$

$$\text{Rate Ratio} = Rate_{exposed}/Rate_{unexposed}$$

Risk ratios and rate ratios are measures of the strength of the association between the exposure and the outcome. How is a risk ratio or rate ratio interpreted? A risk ratio of 1.0 indicates there is no difference in risk between the exposed and unexposed group. A risk ratio greater than 1.0 indicates a positive association, or increased risk for developing the health outcome in the exposed group. A risk ratio of 1.5 indicates that the exposed group has 1.5 times the risk of having the outcome as compared to the unexposed group. Rate ratios can be interpreted the same way but apply to rates rather than risks.

A risk ratio or rate ratio of less than 1.0 indicates a negative association between the exposure and outcome in the exposed group compared to the unexposed group. In this case, the exposure provides a protective effect. For example, a rate ratio of 0.80 where the exposed group received a vaccination for Human Papillomavirus (HPV) indicates that the exposed group (those who received the vaccine) had 0.80 times the rate of HPV compared to those who were unexposed (did not receive the vaccine).

One of the benefits the measure risk difference has over the risk ratio is that it provides the absolute difference in risk, information that is not provided by the ratio of the two. A risk ratio of 2.0 can imply both a doubling of a very small or large risk, and one cannot determine which is the case unless the individual risks are presented.

Odds ratio

Another measure of association is the odds ratio (OR). The formula for the OR is:

$$\text{Odds ratio} = \frac{odds_{exposed}}{odds_{unexposed}}$$

The odds ratio is used in place of the risk ratio or rate ratio in case-control studies. In this type of study, the underlying population at risk for developing the health outcome or disease cannot be determined because individuals are selected as either diseased or non-diseased or as having the health outcome or not having the health outcome. An odds ratio may approximate the risk ratio or rate ratio in instances where the health outcome prevalence is low (less that 10%) and specific sampling techniques are utilized, otherwise there is a tendency for the OR to overestimate the risk ratio or rate ratio.

The odds ratio is interpreted in the same manner as the risk ratio or rate ratio with an OR of 1.0 indicating no association, an OR greater than 1.0 indicating a positive association, and an OR less than 1.0 indicating a negative, or protective association.

The null value

The null value is a number corresponding to no effect, that is, no association between exposure and the health outcome. In epidemiology, the null value for a risk ratio or rate ratio is 1.0, and it is also 1.0 for odds ratios and prevalence ratios (terms you will come across). A risk ratio, rate ratio, odds ratio or prevalence ratio of 1.0 is obtained when, for a risk ratio for example, the risk of disease among the exposed is equal to the risk of disease among the unexposed.

Statistical testing focuses on the *null hypothesis*, which is a statement predicting that there will be no association between exposure and the health outcome (or between the assumed cause and its effect), i.e. that the risk ratio, rate ratio or odds ratio will equal 1.0. If the data obtained from a study provide evidence against the null hypothesis, then this hypothesis can be rejected, and an alternative hypothesis becomes more probable.

For example, a null hypothesis would say that there is no association between children having cigarette smoking mothers and the incidence of asthma in those children. If a study showed that there was a greater incidence of asthma among such children (compared with children of nonsmoking mothers), and that the risk ratio of asthma among children of smoking mothers was 2.5 with a 95% confidence interval of 1.7 to 4.0, we would reject the null hypothesis. The alternative hypothesis could be expressed in two ways: 1) children of smoking mothers will have either a higher or lower incidence of asthma than other children, or 2) children of smoking mothers will only have a higher incidence of asthma. The first alternative hypothesis involves what is called a *"two-sided test"* and is used when we simply have no basis for predicting in which direction from the null value exposure is likely to be associated with the health outcome, or, in other words, whether exposure is likely to be beneficial or harmful. The second alternative hypothesis involves a *"one-sided test"* and is used when we have a reasonable basis to assume that exposure will only be harmful (or if we were studying a therapeutic agent, that it would only be beneficial).

Measures of significance

The p-value

The "p" value is an expression of the probability that the difference between the observed value and the null value has occurred by "chance", or more precisely, has occurred simply because of sampling variability. The smaller the "p" value, the less likely the probability that sampling variability accounts for the difference. Typically, a "p" value less than 0.05, is used as the decision point, meaning that there is less than a 5% probability that the difference between the observed risk ratio, rate ratio, or odds ratio and 1.0 is due to sampling variability. If the "p" value is less than 0.05, the observed risk ratio, rate ratio, or odds ratio is often said to be "statistically significant." However, the use of 0.05 as a cut-point is arbitrary. The exclusive use of "p" values for interpreting results of epidemiologic studies has been strongly discouraged in the more recent texts and literature because research on human health is not conducted to reach a decision point (a "go" or "no go" decision), but rather to obtain evidence that there is reason for concern about certain exposures or lifestyle practices or other factors that may adversely influence the health of the public. Statistical tests of significance, (such as p-values) were developed for industrial quality-control purposes, in order to make a decision whether the manufacture of some item is achieving acceptable quality. We are not making such decisions when we interpret the results of research on human health.

The lower bound of the 95% confidence interval is also often utilized to decide whether a point estimate is statistically significant, i.e. whether the measure of effect (e.g. the ratio 2.5 with a lower bound of 1.8) is statistically different than the null value of 1.0.

Measures of precision

Confidence interval

A *confidence interval* expresses the extent of potential variation in a point estimate (the mean value or risk ratio, rate ratio, or odds ratio). This variation is attributable to the fact that our point estimate of the mean or risk ratio, rate ratio, or odds ratio is based on some sample of the population rather than on the entire population.

For example, from a clinical trial, we might conclude that a new treatment for high blood pressure is 2.5 times as effective as the standard treatment, with a 95% confidence interval of 1.8 to 3.5. 2.5 is the *point estimate* we obtain from this clinical trial. But not all subjects with high blood pressure can be included in any study, thus the estimate of effectiveness, 2.5, is based on a particular sample of people with high blood pressure. If we assume that we could draw other samples of persons from the same underlying population as the one from which subjects were obtained for this study, we would obtain a set of point estimates, not all of which would be exactly 2.5. Some samples would be likely to show an effectiveness less than 2.5, and some greater than 2.5.

The 95% CI is an interval that will contain the true, real (population) parameter value 95% of the time if you repeated the experiment/study. So if we were to repeat the experiment/study, 95 out of 100 intervals would give an interval that contains the true risk ratio, rate ratio or odds ratio value. Remember, that you can only interpret the CI in relation to talking about repeated sampling. Thus we can also say that the new treatment for high blood pressure is 2.5 times as effective as the standard treatment, but this measure could range from a low of 1.8 to a high of 3.5.

The confidence interval also provides information about how precise an estimate is. The tighter, or narrower, the confidence interval, the more precise the estimate. Typically, larger sample sizes will provide a more precise estimate. Estimates with wide confidence intervals should be interpreted with caution.

Other terms

Crude and adjusted values

There are often two types of estimates presented in research articles, *crude* and *adjusted* values. Crude estimates refer to simple measures that do not account for other factors that may be driving the estimate. For instance, a crude death rate would simply be the number of deaths in a calendar year divided by the average population for that year. This may be an appropriate measure in certain circumstances but could become problematic if you want to compare two or more populations that vary on specific factors known to contribute to the death rate. For example, you may want to compare the death rate for two populations, one of which is located in a high air pollution area, to determine if air pollution levels affect the death rate. The high air pollution population may have a higher death rate, but you also determine that it is a much older population. As older individuals are more likely to die, age may be driving the death rate rather than the pollution level. To account for the difference in age distribution of the populations, one would want to calculate an *adjusted* death rate that adjusts for the age structure of the two groups. This would remove the effect of age from the effect of air pollution on mortality.

Adjusted estimates are a means of controlling for confounders or accounting for effect modifiers in analyses. Some factors that are commonly adjusted for include gender, race, socioeconomic status, smoking status, and family history.

Practice Questions

Answers are at the end of this notebook.

1. Based on the following table, calculate the requested measures. Also provide the definition for each measure in one sentence.

a) The risk ratio comparing the exposed and the unexposed study participants

b) The risk difference between the exposed and the unexposed study participants

c) The prevalence of the disease among the entire study sample, assuming the disease is a long-term, chronic disease with no cure and assuming no study participants have died.

	Has disease	Does not have disease	Total
Exposed	651	450	1101
Unexposed	367	145	512
Total	1018	595	1613

2. Interpret the following risk ratios in words.

a) A risk ratio= 1.0 in a study where researchers examined the association between consuming a certain herbal supplement (the exposure) and developing arthritis.
b) A risk ratio= 2.6 in a study where researchers examined the association between ever having texted while driving (the exposure) and being in a car accident.
c) A risk ratio = 0.75 in a study where researchers examined the association between ≥ 30 minutes of daily exercise (the exposure) and heart disease.

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Answers to Practice Questions

1.a) Risk ratio= risk exposed / risk unexposed = (651/1101) / (367/ 512) = 0.82

The risk ratio reflects the ratio of the risk of the disease in the exposed study participants compared with the risk of the disease in the unexposed study participants.

1b) Risk difference = risk exposed - risk unexposed = (651/1101) - (367/512) = -0.13

The risk difference indicates how much excess risk is due to the exposure studied.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers Continued

1c) Prevalence= Total # people with the disease / total # of people in the study population = 1018/1613 = 0.63 Prevalence refers to the proportion of the population studied that has the disease at a given time.

2a) A risk ratio of 1.0 means there is no difference in risk for the health outcome when comparing the exposed and unexposed groups, i.e. the herbal supplement was not associated in any way with the development of arthritis

2b) A risk ratio of 2.6 means there is a positive association, i.e. there is an increased risk for the health outcome among the exposed group when compared with the unexposed group. The exposed group has 2.6 times the risk of having the health outcome when compared with the unexposed group. In this example, the risk ratio of 2.6 means that people who had reported ever texting while driving had 2.6 times the risk of being in a car accident when compared with people who reported never having texted while driving.

2c) A risk ratio of 0.75 means there is an inverse association, i.e. there is a decreased risk for the health outcome among the exposed group when compared with the unexposed group. The exposed group has 0.75 times the risk of having the health outcome when compared with the unexposed group. In this example, the risk ratio of 0.75 means that people who exercised at least 30 minutes per day had 0.75 times the risk of developing heart disease when compared with people who did not exercise at least 30 minutes a day.

ERIC Notebook

Second Edition

Common Statistical Tests and Applications in Epidemiological Literature

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Any individual in the medical field will, at some point, encounter instances when epidemiological methods and statistics will be valuable tools in addressing research questions of interest.

Examples of such questions might include:

- Will treatment with a new anti-hypertensive drug significantly lower mean systolic blood pressure?
- Is a visit with a social worker, in addition to regular medical visits, associated with greater satisfaction of care for cancer patients as compared to those who only have regular medical visits?

There are a number of steps in evaluating data before actually addressing the above questions. These steps include description of your data as well as determining what the appropriate tests are for your data.

Description of data

The type of data one has determines the statistical procedures that are utilized. Data are typically described in a number of ways: by type, distribution, location and variation.

There are three different types of data: nominal, ordinal, and continuous data. Nominal data do not have an established order or rank and contain a finite number of values. Gender and race are examples of nominal data. Ordinal data have a limited number of values between which no other possible values exist. Number of children and stage of disease are good examples of ordinal data. It should be noted that ordinal data do not have to have evenly spaced values as occurs with continuous data, however, there is an implied underlying order. Since both ordinal and nominal data have a finite number of possible values, they are also referred to as discrete data. The last type of data is continuous data which are characterized by having an infinite number of evenly spaced values. Blood pressure and age fall into this category. It should be noted for data collection and analysis that continuous, ordinal, or nominal values can be grouped. Grouped data are often referred to as categorical. Possible categories might include: low, medium, high, or those representing a numerical range.

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

A second characteristic of data description, distribution, refers to the frequencies or probabilities with which values occur within our population. Discrete data are often represented graphically with bar graphs like the one below (Figure 1).

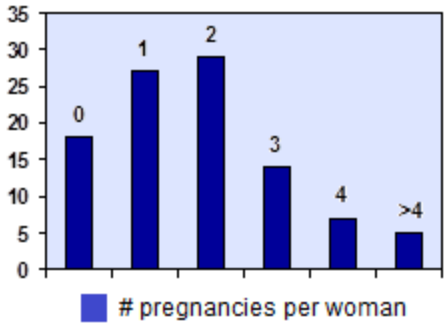


Figure 1. Bar graph

Continuous data are commonly assumed to have a symmetric, bell-shaped curve as shown below (Figure 2). This is known as a Gaussian distribution, the most commonly assumed distribution in statistical analysis.

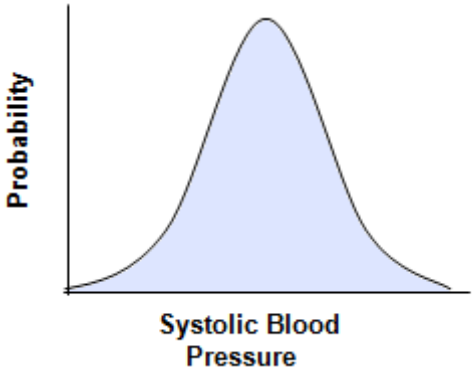


Figure 2. Gaussian distribution

Hypothesis testing

Hypothesis testing, also known as statistical inference or significance testing, involves testing a specified hypothesized condition for a population's parameter. This condition is best described as the null hypothesis. For example, in a clinical trial of a new anti-hypertensive drug, the null hypothesis would state that there is no difference in effect when comparing the new drug to the current standard treatment. Contrary to the null is the alternative hypothesis, which generally defines the possible values for a parameter of interest. For the previous example, the alternative hypothesis is that there is a difference in the mean blood pressure of the standard treatment and new drug group following therapy. The alternative hypothesis might also be described as your "best guess" as to what the values are.

However, in statistical analysis, the null hypothesis is the main interest, and is the one actually being tested. In statistical testing, we assume that the null hypothesis is correct and determine how likely we are to have obtained the sample (or values) we actually obtained in our study under the condition of the null. If we determine that the probability of obtaining the sample we observed is sufficiently small, then we can reject the null hypothesis. Since we are able to reject the null hypothesis, we have evidence that the alternative hypothesis may be true.

On the other hand, if the probability of obtaining our study results is not small, we fail to reject the assumption that the null hypothesis is true. It should be noted that we are not concluding that the null is true. This is a small, but important distinction. A test that fails to reject the null hypothesis should be considered inconclusive. An example will help to illustrate this point.

In a sealed bag, we have 100 blue marbles and 20 red marbles. (This bag is essentially representing the entire population). One individual formulates the null hypothesis that "all the marbles are blue", and the alternative which is "all the marbles are not blue". To test this hypothesis, 10 marbles are sampled from the bag. All ten marbles selected are indeed blue. Thus the individual has failed to reject the null that all the marbles in the bag are blue. However, because all of the marbles were not sampled, you cannot conclude that all the marbles in the bag are blue. (We happen to know this is not true, but it is impossible to know in the real world with populations too large to fully evaluate). If another individual selects 10 marbles from the bag and finds that 8 are blue and 2 are red, we can reject the null hypothesis that all the marbles are blue since we have selected at least one red marble.

Error in statistical testing

Earlier, we indicated that we can reject the null hypothesis if the probability of obtaining a sample like the one observed in our study is sufficiently small. You may ask "What is sufficiently small?" "How small" is determined by how willing we are to reject the null hypothesis when it accurately reflects the population from which it is sampled. This type of error is called a *Type I error*. This error is also commonly called alpha (α). Alpha is the probability of rejecting the null hypothesis when the null is true. This probability is selected by the researcher and is typically set at 0.05. It is important to remember that this is an arbitrary cut-point and should be taken into consideration when making conclusions about the results of the study.

There is a second type of error that can be made during statistical testing. It is known as *Type II error,* which is the probability of not rejecting the null when the alternative hypothesis is indeed true, or in other words, failing to reject the null when the null hypothesis is false. Type II error is commonly known as β. Beta relates to another important parameter in statistical testing which is *power*. Power is equal to (1-β) and is essentially the ability to avoid making a type II error. Like α, power is also defined by the researcher, and is typically set at 0.80. Below is a schematic of the relationships between α, β and power.

Decision	Truth	
	Null True	Null False
Reject Null	α	power
Accept Null		β

Students' T test

This test is most commonly used to test the difference between the means of the dependent variables of two groups. For example, this test would be appropriate if one wanted to evaluate whether or not a new anti-hypertensive drug reduces mean systolic blood pressure.

Example

To evaluate if drug Z reduces mean systolic blood pressure, a randomized clinical trial will be performed where 12 individuals receive drug Z and 8 receive a placebo. The null hypothesis to be tested is that there is no difference in the mean systolic blood pressure of the experimental and placebo groups. The alternative hypothesis is that there is a difference between the means of the two groups. The type I error for your trial will be 5%.

Results

Below is the group assignments and resulting systolic blood pressure (SBP)

Patient	Assignment	Systolic BP
1	Drug Z	100
3	Drug Z	110
5	Drug Z	122
7	Drug Z	109
9	Drug Z	108
11	Drug Z	111
13	Drug Z	118
15	Drug Z	105
17	Drug Z	115
18	Drug Z	119
19	Drug Z	106
20	Drug Z	109
2	Placebo	129
4	Placebo	125
6	Placebo	136
8	Placebo	129
10	Placebo	135
12	Placebo	134
14	Placebo	140
16	Placebo	128

$$\text{mean}_{drug} = \frac{100 + 110 + \ldots + 109}{12} = 111 \text{ mm Hg}$$

$$\text{mean}_{placebo} = \frac{129 + 125 + \ldots + 128}{8} = 132 \text{ mm Hg}$$

$$\text{mean}_{drug} - \text{mean}_{placebo} = -21 \text{ mm Hg}$$

Now that we have determined the difference between means, we need to determine the standard error for that difference which is calculated using the pooled estimate of the variance (σ^2).

The formula for the standard error of the drug Z group is:

$$\sigma^2 drug = \frac{\sum (SBP_{drug} - mean_{drug})^2}{n_{drug} - 1} =$$

$$\sigma^2 drug = \frac{[(100-111)^2 + (110-111)^2 + \ldots + (109-111)^2]}{12-1} = 40.9$$

The standard error for the placebo group is calculated in the same manner substituting the values for the placebo group.

$$\sigma^2 placebo = 25.1$$

Next, we would need to calculate a pooled estimate of the variance using the following equation:

$$\sigma^2_p = \frac{[(n_{drug} - 1)\ \sigma^2_{drug}] + (n_{placebo} - 1)\ \sigma^2_{placebo}]}{(n_{drug} - 1) + (n_{placebo} - 1)} =$$

$$\sigma^2_p = \frac{(11)(40.9) + (7)(25.1)}{11 + 7} = \frac{626}{18} = 34.8$$

The pooled estimate of the variance can then be utilized to calculate the standard error for the difference in means:

$$SE^2 (\text{mean}_{drug} - \text{mean}_{placebo}) = \frac{\sigma^2_p}{n_{drug}} + \frac{\sigma^2_p}{n_{placebo}}$$

$$SE^2 = \frac{34.8}{12} + \frac{34.8}{8} = 7.236$$

$$SE = 2.69$$

Now we are finally ready to test for significant differences in the mean blood pressure of our two groups: (*mean indicates the hypothesized values for the null-generally this quantity would = 0 when there is no difference expected between the drug and placebo groups).

$$t = \frac{(\text{mean}_{drug} - \text{mean}_{placebo}) - (*\text{mean}_{drug} - *\text{mean}_{placebo})}{SE\ (\text{mean}_{drug} - \text{mean}_{placebo})}$$

$$t = \frac{-21 - 0}{2.69} = -7.8 = |-7.8| = 7.8$$

We now compare our calculated value to a table of critical values for the Students' T distribution (found in most basic statistics books). The table also requires that we know the degrees of freedom and the value of a we have selected. Degrees of freedom (df) refers to the amount of information that a sample has in estimating the variance. It is generally the sample size minus one. The df for our calculation is 12 + 8 - 2 = 18 (the sample size for each group - 1). With a two tailed a of 0.05, our value |-7.8| is greater than the critical value from the table (2.101). Thus, we can reject the null hypothesis that there is no difference between mean blood pressure levels, and accept, by elimination, our alternative hypothesis.

Chi-square analysis

What happens if we don't have continuous data, and are faced with categorical data instead? We could turn to chi-square analysis to evaluate if there are significant associations between a given exposure and outcome (the row and column variables in a contingency table). 2 X 2 contingency tables are one of the most common ways to present categorical data, and we can see this in analyzing data that was collected to address the question presented in this notebook.

Is a visit with a social worker, in addition to regular medical visits, associated with greater satisfaction of care for cancer patients as compared to those who only have regular medical visits?

Below is a generic 2 X 2 table representing the data. It is important to note the set-up of the table, as cell "a" generally represents the group of interest (diseased and exposed) and cell d represents the referent group (no disease and unexposed).

	Row value (often disease or health outcome)		
Column Value (often Exposure)	1	0	**Total**
1	*a*	*b*	*a + b*
0	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n*

Here we have the contingency table with data from our trial:

	Greater Satisfaction?		
Social Worker Visit?	Yes	No	**Total**
Yes	64	46	110
No	36	54	90
Total	100	100	200

In chi-square analysis we are testing the null hypothesis that there is no association between a social worker visit and a greater satisfaction with care.

Generally, in evaluating this type of data, it is important for each of the individual cells to have large values, (i.e. greater than 5 or 10 each), If these conditions are not met, a special type of chi-square analysis is conducted called the Fisher's exact test. This will not be discussed in this notebook.

To calculate the chi-square statistic (χ^2):

$$\chi^2 = \sum \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

with *i* representing the frequency in a particular cell of the 2 X 2 table. Below is the calculation for the frequencies that are **expected** in each cell.

	Row value		
Column Value	1	2	**Total**
1	$\frac{(a+b)(a+c)}{n}$	$\frac{(a+b)(b+d)}{n}$	$a + b$
2	$\frac{(c+d)(a+c)}{n}$	$\frac{(c+d)(b+c)}{n}$	$c + d$
Total	$a + c$	$b + d$	n

Thus, we now have a table that has both the actual and expected (in parentheses) values:

	Greater Satisfaction?		
Social Worker Visit?	Yes	No	**Total**
Yes	64 (55)	46 (55)	110
No	36 (45)	54 (45)	90
Total	100	100	200

With this information, we can now calculate the χ^2 statistic:

$$\chi^2 = \sum \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$$\chi^2 = \frac{(64-55)^2}{55} + \frac{(46-55)^2}{55} + \frac{(36-45)^2}{45} + \frac{(54-45)^2}{45}$$

$$\chi^2 = 6.545$$

The chi-square statistic for these data has approximately 1 degree of freedom, an α of 0.05, and it is compared to the critical values on standard Chi-square table. Note that the degrees of freedom would increase as the number of rows and columns of our tables increases (for instance a 3 X 4 table). Since our calculated value ($\chi 2 = 6.545$) is greater than the critical value (3.841), we can once again reject the null hypothesis that there is no association between the exposure and the outcome of interest, and conclude that in this case seeing a social worker is significantly associated with a greater satisfaction with care.

Important notes

It is important to remember that the statistical tests and examples presented here are only an elementary presentation of the large scope of situations that can be addressed by these data. The intention of this notebook is to provide a basic understanding of the underlying principles of these statistical tests rather than implying that what has been presented is appropriate for every situation.

Further information about these statistical tests and other applications can be found in the following references:

Statistical First Aid: Interpretation of Health Research Data by Robert P Hirsch and Richard K. Riegelman. Blackwell Scientific Publications, Cambridge, MA 1992.

Categorical Data Analysis, Using the SAS System by ME Stokes, CS Davis, and GG Koch. SAS Institute Inc., Cary, NC, 2001.

Practice Questions

Answers are at the end of this notebook

Researchers are conducting a study of the association between working in a noisy job environment and hearing loss. The researchers' null hypothesis is that there is no difference in hearing loss between people who work in a noisy job environment compared with people who work in a quiet job environment. The researchers' alternative hypothesis is that there is a difference in hearing loss between people who work in a noisy job environment compared with people who work in a quiet job environment. The researchers decided to set their alpha level at 0.05. The researchers' analysis results show a p-value of 0.0003 (please note that for the purposes of this question you are being provided with just the p-value from the study when in reality a study analysis is much more complex).

1) True or false: The alpha level of 0.05 is an arbitrary value.

2) True or false: Based on the results, the researchers can conclude their null hypothesis is true.

3) True or false: Based on these results, the researchers should reject the assumption that their null hypothesis is true.

4) True or false: An alpha level of 0.05 means there is a 0.05 percent chance that the researchers will incorrectly reject the null hypothesis.

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1) True or false: The alpha level of 0.05 is an arbitrary value.

Answer: True

This statement is true. The level of alpha is often set at 0.05, however, this choice is arbitrary and researchers may choose a different value.

2) True or false: Based on the results, the researchers can conclude their null hypothesis is true.

Answer: False

This statement is false. Researchers should never conclude that their null hypothesis is true. It is possible to conclude that we should fail to reject the assumption that the null hypothesis is true but this is not the same as concluding that the null hypothesis is actually true.

3) True or false: Based on these results, the researchers should reject the assumption that their null hypothesis is true.

Answer: True

This statement is true. The p-value was 0.0003 which is less than the alpha of 0.05. The researchers should reject their null hypothesis that there is no difference in hearing loss diagnosis between people who work in a noisy job environment compared with people who work in a quiet job environment.

4) True or false: An alpha level of 0.05 means there is a 0.05 percent chance that the researchers will incorrectly reject the null hypothesis.

Answer: False

This statement is false. An alpha level of 0.05 means there is a 5% chance the researchers will incorrectly reject the null hypothesis.

ERIC Notebook

Second Edition

Confounding Bias, Part I

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Confounding is one type of systematic error that can occur in epidemiologic studies. Other types of systematic error such as information bias or selection bias are discussed in other ERIC notebook issues.

Confounding is an important concept in epidemiology, because, if present, it can cause an over- or under-estimate of the observed association between exposure and health outcome. The distortion introduced by a confounding factor can be large, and it can even change the apparent direction of an effect. However, unlike selection and information bias, it can be adjusted for in the analysis.

What is confounding?

Confounding is the distortion of the association between an exposure and health outcome by an extraneous, third variable called a confounder. Since the exposure of interest is rarely the only factor that differs between exposed and unexposed groups, and that also affects the health outcome or disease frequency, confounding is a common occurrence in etiologic studies.

Confounding is also a form a bias. Confounding is a bias because it can result in a distortion in the measure of association between an exposure and health outcome.

Confounding may be present in any study design (i.e., cohort, case-control, observational, ecological), primarily because it's not a result of the study design. However, of all study designs, ecological studies are the most susceptible to confounding, because it is more difficult to control for confounders at the aggregate level of data. In all other cases, as long as there are available data on potential confounders, they can be adjusted for during analysis.

Confounding should be of concern under the following conditions:

1. Evaluating an exposure-health outcome association.
2. Quantifying the degree of association between an exposure and health outcome. For example, you might want to quantify how being overweight increases the risk of cardiovascular disease (CVD). If you were concerned about age as a confounder, you would “control for” the effect of age in your statistical modeling.

In one study, the rate ratio might change from 4.0 to 3.7 when controlling for age, whereas in another study, a rate ratio of 4 may change to 1.2 after controlling for age.

3. Multiple causal pathways may lead to the health outcome. If there is only one way to contract the health outcome or disease, confounding cannot occur. This criterion is almost always met as health outcomes can inevitably be caused by different agents, different transmission routes, or different biological or social mechanisms.

A few examples of research questions in which you would want to consider confounding are listed below:

1. Does being overweight increase the risk of coronary heart disease (CHD) -- independently of cholesterol, hypertension, and diabetes?
2. Does tobacco advertising entice adolescents to experiment with tobacco independently of whether or not their parents smoke?

Assessing confounding

Each potential confounder has to meet two criteria before they can be confounders: *Criterion 1* is that the potential confounder must be a known risk factor for the health outcome or disease.

Broadly speaking, a risk factor is any variable that is:

1. Already known to be "causally related" to the health outcome or disease (though not necessarily a direct cause) AND
2. Antecedent to the health outcome or disease on the basis of substantive knowledge or theory, and/or on previous research findings.

The confounding factor must be predictive of the health outcome or disease occurrence apart from its association with exposure; that is, among unexposed (reference) individuals, the potentially confounding factor should be related to the health outcome or disease.

With an epidemiological data set, one can calculate whether or not a potential confounder is a risk factor using the following mathematical formula:

Criterion 1 for confounding: mathematical formula

Criterion 1 for confounding is the following: among the unexposed, there should be an association between the confounder and the health outcome.

To convert this to a mathematical equation, the first thing to realize is that Criterion 1 involves calculating a measure of association ("there should be an association between the confounder and the health outcome"). Examples of measures of association are: risk ratios, rate ratios, odds ratios, and risk differences – the type of measure depends on the type of data available, and the scale on which the measure of association is assessed (additive or multiplicative scale). This measure of association will be calculated among the unexposed population only.

For a prospective cohort study where we want to measure the association on a multiplicative scale, we will calculate the following rate ratio (RR):

$$RR_{CD/E-} = \text{risk ratio confounder in unexposed} = \frac{\text{Rate of new cases among population A}}{\text{Rate of new cases among population B}}$$

where the rate of new cases = the number of new cases divided by the total number of susceptible individuals. Population A is comprised of all individuals who have the confounder (C+) but who are unexposed (E-), and population B is comprised of all individuals who don't have the confounder (C-) or the exposure (E-).

For a case-control study using odds ratios (OR), the formula for Criterion 1 is:

ORCD/E- = odds ratio confounder in unexposed

$$\frac{\text{Odds that cases have confounder among population F}}{\text{Odds that controls have confounder among population F}}$$

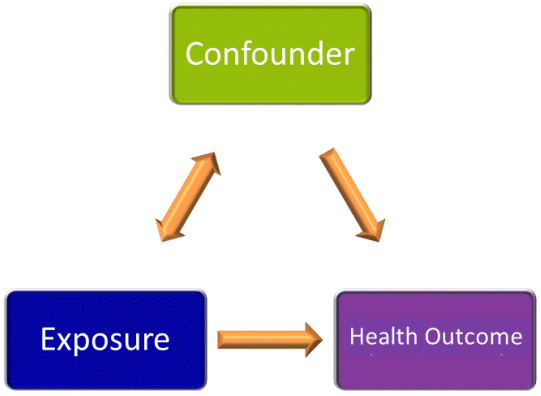
where the odds that the cases have the confounder = the number of cases with the confounder (C+) divided by the number of cases without the confounder (C-) and where population F is comprised of all individuals who are not exposed (E-).

Now that the risk ratio, rate ratio, or odds ratio for the association between the confounder and health outcome among the unexposed has been calculated, how is it interpreted?

For the confounder to be a risk factor, the measure of association has to be greater than 1 (for a harmful association), or less than 1 (for a protective association).

Age and smoking status, for example, are widely considered to be risk factors for lung cancer, even though the mechanisms by which both variables are determinants of this disease are not well understood. On the other hand, race is not considered to be a risk factor for lung cancer. Unnecessary adjustment of variables that are not confounders can lower precision and may even introduce bias into the estimate of effect.

Criterion 2 is that the potential confounder must be associated with the main exposure, but not as a result of the exposure. In other words, all potential confounders should be working independently and not as part of the proposed exposure-health outcome pathway. One can calculate whether or not a potential confounder is associated with the main exposure using a mathematical formula.



Criterion 2 for confounding: mathematical formula

Criterion 2 for confounding is the following: the distribution of the confounding variable differs between exposed and unexposed groups.

To convert this to a mathematical equation, the first thing to realize is that Criterion 2 involves calculating a measure of association.

For a prospective cohort study, we will calculate the following risk ratio:

RR_{EC}=

$$\frac{\text{\% individuals with confounder (C+) among Population A}}{\text{\% individuals with confounder (C+) among Population B}}$$

where Population A will be comprised of all individuals who are exposed (E+), and where Population B will be comprised of all individuals who are unexposed (E-).

For a case-control study using odds ratios (OR) the formula for Criterion 2 is:

OR_{EC}=

$$\frac{\text{Odds of controls having the confounder (C+) among Population A}}{\text{Odds of controls having the confounder (C+) among Population B}}$$

where odds of controls having the confounder (C+) = number of controls having the confounder (C+) divided by the number of controls not having the confounder (C-). Population A is comprised of all individuals who are exposed (E+), and population B is comprised of all individuals who

are unexposed (E-). Note the additional inclusion criteria for case-control studies: the individuals included in this calculation must include only those who have the potential to be cases (the control group).

Now that the risk ratio, rate ratio, or odds ratio for the association between the confounder and exposure has been calculated, how is it interpreted? For the confounder associated with the exposure, this association has to be greater than 1 (for a harmful association) or less than 1 (for a protective association).

To decide whether a variable is working independently of the association of interest, there must be a biological or social mechanism to causally link the exposure of interest to the disease or health outcome. Such decisions should be made on the basis of the best available information, including non-epidemiological (i.e., clinical, sociological, psychological, or basic science) data. This criterion is obviously satisfied if the confounding factor precedes the exposure and health outcome or disease.

For instance, if interested in assessing the association between physical inactivity and cardiovascular disease (CVD), body weight should not be controlled for if being overweight may be an intermediary step in the causal pathway between physical inactivity and CVD.

In contrast, if the proposed causal pathway is independent of body weight, then body weight can be considered a potential confounder. If intervening variables are controlled for in the analysis, it may reduce or eliminate any indications in the data of a true association between disease and exposure.

ERIC Notebook *Confounding Bias Part II and Effect Measure Modification,* discuss control of confounders in epidemiological studies.

Terminology

Confounding bias: A systematic distortion in the measure of association between exposure and the health outcome caused by mixing the effect of the exposure of primary interest with extraneous risk factors.

Practice Questions

Answers are at the end of this notebook

Researchers have conducted a cohort study in country A to examine the association between a diet high in fat and the risk of colon cancer. The researchers believe that vitamin use may be a confounder. Use the 2x2 tables below to determine if vitamin use is a confounder in the high fat diet-colon cancer association.

	Colon cancer	No colon cancer	Total
Exposed to a high fat diet	254	2220	2474
Not exposed to a high fat diet	150	1500	1650

Among people exposed to a high fat diet (n=2474):

	Colon cancer	No colon cancer	Total
Takes daily vitamin	150	1830	1980
Does not take daily vitamin	104	390	494

Among people <u>not</u> exposed to a high fat diet (n=1650):

	Colon cancer	No colon cancer	Total
Takes daily vitamin	50	800	850
Does not take daily vitamin	100	700	800

1) Is vitamin use an independent risk factor or protective factor for colon cancer?

2) Is vitamin use differentially distributed between the high fat diet and low fat diet groups?

3) Compare the crude risk ratio with the risk ratios stratified by vitamin use.

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the

Answers to Practice Questions

1) Risk ratio of vitamin users getting colon cancer among the non-exposed group: (50/850) / (100/800)= 0.47

A risk ratio of 0.47 shows that vitamin use is a moderate inverse predictor of colon cancer. In this study population, vitamin use was protective for colon cancer.

2 Among people who eat a high fat diet there are 1980/2474= 80% vitamin users

Among people who do not eat a high fat diet there are 850/1650= 52% vitamin users

So vitamin use is differentially distributed among the high fat and low fat diet exposure groups.

3) The crude risk ratio (not stratified by vitamin use) is the risk of colon cancer from high fat diet exposure / the risk of colon cancer from low fat diet exposure. Crude risk ratio = (254/2474) / (150/1650) = 1.13

The risk ratio for colon cancer among vitamin users with a high fat diet is:
Risk ratio = (150/1980) / (50/850)= 1.29

The risk ratio for colon cancer among non-vitamin users with a high fat diet is:

Risk ratio = (104/494) / (100/800) =1.68

The crude risk ratio of 1.13 and the vitamin-specific risk ratio of 0.47 (from question 1) are not in between the stratified risk ratios, they are both lower than the stratified risk ratios. Thus, the crude risk ratio is confounded by vitamin use.

ERIC Notebook

Second Edition

Confounding Bias, Part II and Effect Measure Modification

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Confounding is one type of systematic error that can occur in epidemiologic studies, and is a distortion of the association between an exposure and health outcome by an extraneous, third variable

Effect measure modification (EMM) is when a measure of association, such as a risk ratio, changes over values of some other variable. In contrast to confounding which is a distortion, EMM is of scientific interest ,answers a research question, and can help identify susceptible or vulnerable populations. Are children with poor nutrition, exposed to water pollution, at higher risk of gastrointestinal symptoms compared with children with good nutrition who are exposed to water pollution?

Both confounding and EMM can be considered in the design phase of research or in the analysis phase. We will discuss both in this notebook.

Calculating and adjusting for confounding

The previous issue of ERIC Notebook, "Confounding Bias, Part I", discussed two criteria for identifying potential confounders in a study. Once potential confounders have been identified, the next step is to evaluate if and how much the confounders bias the study results. To do this, results where confounding is ignored, the *"crude"* measure of association, are compared to results that have been corrected for distortions due to confounding, the *"adjusted"* measure of association.

This methodology makes 2 assumptions:

First, the data are obtained by simple random sampling rather than by some more restrictive subject selection procedure, like matching.

The second assumption is that the exposure, health outcome, and confounder variables are all dichotomous (i.e., having only two strata). If the variables are in a continuous format, they can either be dichotomized, or they must be adjusted for simultaneously to calculate the true measure of association.

Methods to calculate adjusted measures of associations differ by the need to control each confounder individually or all confounders simultaneously.

Before calculating an adjusted measure of association using stratified analyses, one must first assess the presence of *effect measure modification* (EMM). When effect measure modification is present, it can be difficult to ascertain whether or not confounding is occurring.

Two examples of EMM are:

- A breast cancer education program (the exposure) that is much more effective in reducing breast cancer in rural areas than urban areas. Here, the area (rural or urban) is an effect measure modifier.
- The finding that a reduction in regional public transportation services (the exposure) affects individuals with little or no access to a car much more than those individuals with access to a car. In this example, having access to a car is the effect measure modifier.

What is effect measure modification?

When estimates of an exposure-health outcome relationship stratified by a confounder are sufficiently different from one another (i.e., risk ratio level 1=4.0 and risk ratio level 2 =0.2), they suggest that two different exposure-health outcome relationships may be operating, one in each level of the confounder.

EMM is different from confounding, where instead of "competing" with the exposure of interest in explaining the etiology of a health outcome or disease, the effect measure modifier identifies subpopulations that are particularly susceptible to the exposure of interest.

Is effect measure modification present?

To calculate whether EMM may be occurring in the study, first calculate three measures of association:

A = The overall, crude measure of association of the exposure-health outcome

B1 = The measure of the exposure-health outcome association among all study participants who have a history of the confounding variable (C+)

B2 = The measure of the exposure-health outcome association among all study participants who do not have a history of the confounding variable (C-)

Use the Figure 1 below as a guide on how to interpret the meaning of these three measures of associations.

As a general rule, if B1 and B2 are basically equal in value, but different from A, then confounding is present and EMM is not present. EMM is present when B1 and B2 are different from one another, and at least one (B1 or B2) is different from A. Both EMM and confounding can occur simultaneously.

If stratified analysis is used to adjust for EMM, confounders should be addressed using more complex statistical techniques, as stratifying results on more than

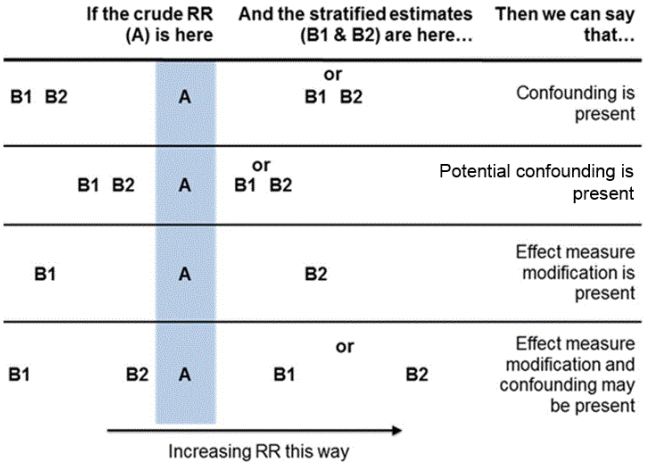


Figure 1

Risk Ratio Example

one variable splits the sample into significantly smaller sample sizes and limits generalizability.

Calculating adjusted summary estimates

If no EMM is present, then stratum-specific estimated effects can be pooled to form a summary estimate of effect across strata. This summary estimate represents an adjusted risk ratio (a risk ratio adjusted for confounding).

Although there are many ways to calculate the adjusted risk ratio, the Mantel Haenszel procedure, is the most common pooling procedure.

Calculating adjusted measures when all confounders are assessed simultaneously

Simultaneous control of two or more variables can give different (and potentially more interesting) results from those obtained by controlling for each variable separately. Simultaneous control of confounders better emulates the natural environment where exposures, diseases, and confounders of interest are found, than does individual control of confounders.

Simultaneous control of several confounders to calculate adjusted measures is done through mathematical modeling.

To control for confounding using mathematical modeling, simply include the confounding variables as independent variables in the model. The simplicity of this method of adjustment for confounding is one of the attractive features of using mathematical models in epidemiology.

Although many types of mathematical models are available, there is generally only one type of model that is appropriate for the goals of a specific data analysis and for the type of data available. The most common mathematical model used in epidemiology is logistic regression.

This model has the general format of

$$y = a + b_{1x} + b_{2z2} + \ldots + b_{izi}$$

Individual-level data need to be provided on:

1. y: the health outcome in a dichotomous format
2. x: the exposure
3. z2 to zi: the confounders

With this information and statistical analysis software, researchers can then calculate the following:

1. a: the y-axis intercept
2. b1: the coefficient for the exposure variable
3. b2 to bi: the coefficients for each confounder that is controlled for in the model.

These coefficients (except for a) are very useful, as they can be transformed into odds ratios. The odds ratio obtained from b1 of this model is an interpretable measure of association describing the relationship between the exposure (x) and the health outcome (y) after adjustment for confounding variables (z_2 to z_i).

The biggest disadvantages to using mathematical models are the assumptions that must be met by the dataset in order to use them – often the data may not conform to all assumptions. It is advisable to do regression diagnostics sometime during the data analysis stage to check these assumptions.

Is adjustment for confounding necessary?

If the adjusted effects are markedly different from the crude effect (typically a 10 or 15% change from crude to adjusted), then confounding is present and should be controlled for. Researchers should report the cut-off they used for their analysis (e.g. 5%, 10%, or 50%) in the selection of confounders for adjustment.

If the adjustment of confounding variables changes the results only slightly (less than 10%), then the tendency would be to ignore this influence, since the more variables controlled for, the less precise (the wider the confidence intervals) the study results will be. The benefits of ignoring the minor confounders would outweigh the costs.

Also consider whether it is important to control for potential confounders such as age, simply because many

readers would not trust results that are not adjusted for age. This distrust stems from knowledge that age is strongly related to disease and mortality rates (similar comments would apply to sex).

Control of confounding

In the analysis phase:

Once data have been collected, there are two options for control of confounding: Stratified analysis or mathematical modeling. Both methods were described above when calculating the effect of confounding on the measure of association. Briefly, stratified analysis pools the measure of association calculated in each strata of the confounder into one summary estimate. Mathematical modeling uses a more complex approach and makes more assumptions than stratified analysis.

In the design phase:

Some confounders should be controlled for in the study design stage of a study, rather than in the analysis stage. It may be necessary to do this if the confounder is very strong and when the anticipated sample size will be large enough to deal with it in the design stage. Some study designs are more favorable for controlling for confounding than others.

Restriction, matching, and randomization are common techniques used to minimize confounding in the design phase. These techniques are not exclusive to one another. Several different control methods may be used at once.

Restriction

Confounding can be controlled for by restricting the study population to those who are unexposed to one or more confounding variables. An example of restriction is to restrict a study population to nonsmokers when studying the association of environmental radon with lung cancer. Restriction is ideal when the exposure-health outcome relationship has strong confounders because it can be an efficient, convenient, inexpensive, and straight-forward method of controlling for confounding. However, the restricted variable, for instance smoking in the given example, cannot be assessed for confounding. Restriction may not always be logistically feasible because the sample size of available study participants is decreased, sometimes to the point that a study cannot be done.

Example

For instance, a group of 30 HIV-positive skydivers of varying age (20% twenty-something, 30% thirty-something, 50% senior citizens) has been identified with which to study behavioral risk factors for HIV infection, independent of age. Therefore, the control group should be 30 HIV-negative skydivers with the same age distribution as the group of HIV-positive skydivers (20% twenty-something, 30% thirty-something, 50% senior citizens). This would be accomplished through matching the controls to the cases by age, by selecting only HIV-negative skydivers who contribute to the pre-determined age distribution.

Matching

Confounding can also be controlled through matching on the confounder variable(s). Matching involves constraining the control group (for case-control studies) or the unexposed group (for cohort studies) such that the distribution of the confounding variable(s) within these groups are similar (or identical) to the corresponding distribution within the index group (the case group for case -control studies or the exposed group for cohort studies). Matching can be viewed as imposing a "partial restriction" on the values of the confounding variables, since only the control or unexposed group is restricted.

Analysis of matched data requires special consideration, because the control, or unexposed, group is not a random sample of study participants; they should be considered to be a biased sample. Techniques for analyzing matched data include conducting the data analysis separately for each level of the confounder (stratified analysis) and using conditional logistic regression.

When considering matching, consider four factors:

1. Precision (generally increased with matching)
2. Cost (generally lowered with matching, because a smaller sample size is needed)

3. Feasibility (can be increased with matching)
4. Flexibility in deciding whether to match

Also keep in mind that variables matched cannot be assessed for confounding.

Randomization

Randomization is an ideal method for controlling for confounding because this method can control both known and unknown confounders. However, because randomization requires that the exposure status of individuals be assigned to study participants, observational study designs such as cross-sectional, cohort, case-control and ecological studies cannot use randomization to control for confounding. For controlled clinical trials however, randomization is a common method to control for confounding.

Use of randomization to control for confounding presumes that random classification of individuals into groups will produce groups that have an equal (or similar) distribution of confounders. For example, the theory of randomization says that given two randomly selected groups of students, each group will have an equal percentage of females, an equal percentage of individuals with white-colored tops, an equal percentage of brown-eyed individuals, and so forth. Thus, randomization, if done correctly, will produce homogeneous groups of individuals. When an exposure is applied to one of these homogeneous groups, but not to the other, the only difference between the two groups is their exposure status. In this situation, confounding, the unequal distribution of a risk factor between exposed and non-exposed groups, cannot occur.

Terminology

Effect measure modification: a variation in the magnitude of a measure of exposure effect across levels of another variable

Randomization: random assignment of subjects to exposure categories

Matching: the selection of controls, or unexposed subjects, that are identical, or nearly so, to the cases, or exposed subjects, with respect to the distribution of one or more potentially confounding factors

From: Modern Epidemiology, Rothman KJ and Greenland S, 1998

The key to proper control of confounding through randomization is having a sufficiently large sample size in each randomized group. Rothman and Greenland (1998) state that having at least 50 subjects, preferably 100 or more, will usually assure that potential confounders are equally distributed among each study group.

Practice Questions
Answers are at the end of this notebook

1) There is a vaccination campaign in effect in some counties of a state to educate parents about childhood diseases and recommended vaccines for their children. Researchers find that the vaccination campaign seems to be effective. Children who live in the counties with the vaccination campaign are 4 times as likely to be vaccinated according to recommendations when compared with children who live in counties without the vaccination campaign. However, even looking only within counties with the vaccination campaign, researchers find that children of families with incomes above the poverty level are twice as likely to vaccinate their children when compared with children of families with incomes below the poverty level. Which of the following can be determined from this example?

Choose all that apply:

a) Family income appears to be a confounder
b) Family income appears to be an effect measure modifier
c) The vaccination campaign appears to be effective
d) The vaccination campaign is the study outcome
e) The vaccination campaign is the study exposure

2) Researchers conducted a cohort study in country B to examine the association between a diet high in fat and the risk of colon cancer. The researchers believe that vitamin use may be a confounder but they first need to examine whether or not vitamin use is an effect measure modifier. Use the 2x2 tables below to determine if vitamin use is an

effect measure modifier or a confounder in the high fat diet- colon cancer association.

	Colon cancer	No colon cancer	Total
Exposed to a high fat diet	254	2220	2474
Not exposed to a high fat diet	98	2300	2398

Among people exposed to a high fat diet:

	Colon cancer	No colon can-	Total
Takes daily	150	1830	1980
Does not take	104	390	494

Among people not exposed to a high fat diet:

	Colon cancer	No colon cancer	Total
Takes daily vitamin	80	1500	1580
Does not take daily vitamin	18	800	818

a) Is vitamin use an independent risk factor for colon cancer?

b) Is vitamin use differentially distributed between the high fat diet and low fat diet groups?

c) Compare the crude risk ratio with the risk ratios stratified by vitamin use.

References

Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. Belmont, CA: Lifetime Learning Publications, 1982.

Miettinen OS, Cook EF. Confounding: essence and detection. American Journal of Epidemiology 114(4):593-603, 1981 Oct.

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

Rothman KJ. Introduction to Epidemiology. New York, NY: Oxford University Press, 2002

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. Answer choices b, c and e are correct. In this example, family income appears to be an effect measure modifier because the effect of the vaccination campaign differs markedly between the 2 sub-populations (family income above versus below poverty level). The vaccination campaign does appear to be effective since children in counties with the vaccination campaign were 4 times as likely to be vaccinated as children in counties without the vaccination campaign. In this example, the researchers are studying the outcome of whether or not children are vaccinated according to recommendations. The vaccination campaign is the exposure.

2.

a) Is vitamin use an independent risk factor for colon cancer?

Answer:
Risk ratio of vitamin users getting colon cancer among the non-exposed group:
(80/1580) / (18/818)=2.3

Yes, a risk ratio of 2.3 shows that vitamin use is a moderate predictor of colon cancer.

b) Is vitamin use differentially distributed between the high fat diet and low fat diet groups?

Answer:
Among people who eat a high fat diet there are
1980/2474= 80% vitamin users
Among people who do not eat a high fat diet there are
1580/2398= 66 % vitamin users
So vitamin use is differentially distributed among the high fat and low fat diet exposure groups.

c) Compare the crude risk ratio with the risk ratios stratified by vitamin use.

Answer:
The crude risk ratio (not stratified by vitamin use) is the risk of colon cancer from high fat diet exposure / the risk of colon cancer from low fat diet exposure.
Crude risk ratio = (254/2474) / (98/2398) = 2.51

The risk ratio for colon cancer among vitamin users with a high fat diet is:
Risk ratio = (150/1980) / (80/1580)= 1.50

The risk ratio for colon cancer among non-vitamin users with a high fat diet is:
Risk ratio = (104/494) / (18/818) = 9.6

The crude risk ratio of 2.51 and the vitamin-specific risk ratio of 2.3 (from question 1) are both in between the stratified risk ratios. The 2 stratified risk ratios differ greatly from one another (9.6 for non-vitamin users versus 1.5 for vitamin users). This example shows the presence of effect measure modification.

ERIC Notebook

Second Edition

Cross-sectional Studies

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Like cohort studies, cross-sectional studies conceptually begin with a population base. But unlike cohort studies, in cross-sectional studies we do not follow individuals over time. Instead, we only look at the prevalence of disease and/or exposure at one moment in time. These studies take a "snapshot" of the proportion of individuals in the population that are, for example, diseased and non-diseased at one point in time. Other health outcomes besides diseases may also be studied. Cross-sectional studies also differ from cohort studies in the populations that are studied. Cohort studies begin by selecting a population of persons who are at risk of for a specific disease or health outcome; cross-sectional studies begin by selecting a sample population and then obtaining data to classify all individuals in the sample as either having or not having the health outcome.

	Cohort	Cross-sectional
Study group	Population-at-risk	Entire population (or a sample)
Common Measures	Risks and Rates	Prevalence

Ways to use cross-sectional studies

Cross-sectional studies are used both descriptively and analytically. *Descriptive cross-sectional studies* simply characterize the prevalence of a health outcome in a specified population. Prevalence can be assessed at either one point in time (*point prevalence*) or over a defined period of time (*period prevalence*). Period prevalence is required when it takes time to accumulate sufficient information on a disease in a population, e.g. what proportion of persons served by a public health clinic over a year have hypertension. These prevalence measures are commonly used in public health; often the point or period aspect is not specified.

In *analytical cross-sectional studies,* data on the prevalence of both exposure and a health outcome are obtained for the purpose of comparing health outcome differences between exposed and unexposed.

Analytical studies attempt to describe the prevalence of, for example, disease or non-disease by first beginning with a population base. These studies differ from solely descriptive cross-sectional studies in that they compare the proportion of

exposed persons who are diseased (a/(a+b)) with the proportion of non-exposed persons who are diseased (c/(c+d)).

Calculating prevalence

The prevalence of a health outcome is simply the proportion of individuals with the health outcome in a population.

Prevalence = cases / total population

For the following example, two different sub-measures of prevalence can be calculated: the prevalence of coronary heart disease (CHD) among the exposed (people who are not active) and the prevalence of CHD among the unexposed.

Example:

	Present		Absent CHD		Total
Not active	50	a	b	200	250
Active	50	c	d	700	750
Total	100			900	1000

P_1= a/a+b= 50/250 = 20.0% prevalence of CHD among people who are not active.

P_0= c/c+d = 50/750 = 6.7% prevalence of CHD among people who are active.

The prevalence odds ratio

The prevalence odds ratio (POR) is calculated in the same manner as the odds ratio.

POR = ad / bc

The prevalence ratio

The prevalence ratio (PR) is analogous to the risk ratio (RR) of cohort studies. The denominators for both ratios are fixed populations – fixed at the start of the study in the case of a cohort study, and fixed at the point or period of time for the case-control study. The PR is similar to a RR when the outcome occurs over a short period of time. For example, one would calculate a prevalence ratio for an acute outbreak of tuberculosis in a prison population. This is in contrast to calculating the overall prevalence of positive tuberculin skin tests among the prisoners.

The prevalence ratio can also be calculated from the information on CHD and physical activity:

PR = (a/N1) / (c/N0)

PR= (50/250) / (50/750) = 3.0

In this case, a prevalence ratio of 3.0 can be interpreted to mean that the people who are not physically active are 3 times as likely as those who are physically active to have CHD.

Note that it is preferable to calculate the prevalence odds ratio when the period for being at risk of developing the outcome extends over a considerable time (months to years) as it does in this example. **POR vs. PR**

For chronic disease studies or studies of long-lasting risk factors, POR is the preferred measure of association in cross-sectional studies. For acute disease studies, PR is the preferred measure of association. If the prevalence of disease is low, i.e. 10% or less in exposed and unexposed populations, POR = PR. Since cross-sectional studies are particularly useful for investigating chronic diseases (e.g. prevalence of AIDS) where the onset of disease is difficult to determine, or for studying long lasting risk factors (such as smoking, hypertension, and high fat diets), the prevalence odds ratio will generally be the preferred measure of association.

Limitations of cross-sectional studies to evaluate risk

Recall that, under steady conditions, the prevalence of disease is influenced both by incidence and duration of disease (or survival with disease).

Prevalence = Rate x Average Duration of Disease

Persons who survive longer with a disease will have a higher probability of being counted in the numerator of a prevalence proportion. Short-term survivors will be less likely to be counted as a case. Incidence is influenced only by exposure, whereas prevalence is influenced both by exposure and duration of disease.

If exposure influences survival time, then the POR or PR will not provide a valid estimate of the risk ratio or rate ratio. Thus, the interpretation of the POR or PR is subject to survival bias.

Even if incidence remains constant, either an improvement in disease treatment (that results in higher cure rates) or increased lethality (resulting in a higher case fatality rate) will result in decreased prevalence. The disease itself or the threat of developing the disease may cause outmigration of cases from an environment perceived as causing disease, e.g. workers affected by toxic exposures in a plant may quit, while more resistant workers will stay. This selective migration can bias measures of prevalence.

Other problems with interpretation of cross-sectional studies

Cross-sectional studies as well as case-control studies are affected by the *antecedent-consequent bias*, similar to the chicken and egg question (i.e. "which came first?"). This bias occurs when it cannot be determined that exposure preceded disease, since both are ascertained at the same time (unlike cohort studies or clinical trials). Antecedent-consequent bias does not affect cohort studies because subjects in cohort studies are selected for study because they are disease-free. Exposure is actually observed to precede disease only in a cohort design, including randomized trials.

Uses of cross-sectional studies

Descriptive studies are an important method to evaluate the proportion of a population with disease or with risk factors for disease, such as the prevalence of asthma in children or the prevalence of elevated blood lead in toddlers.

Descriptive cross-sectional studies are widely used to estimate the occurrence of risk factors in segments of the population characterized by age, sex, race or socioeconomic status (SES). National examples of cross-sectional studies of great importance are the *decennial census* and the National Health and Nutrition Surveys (NHANES). Opinion polls and political polls are basically cross-sectional studies. Surveillance of changes in smoking habits or of other behavioral risk factors are sequential cross-sectional studies. The US National Health and Nutrition Examination Survey (NHANES) is one such example Similarly, surveillance of long lasting diseases such as AIDS is cross-sectional. Descriptive cross-sectional studies are useful for planning or administering preventive or health care services, surveillance programs, and surveys and polls.

Descriptive/analytical cross-sectional studies are useful for establishing preliminary evidence for a causal relationship. These studies are also useful for examining the association between exposure and disease onset for chronic diseases where researchers lack information on time of onset. Examples might include diet and arthritis, smoking and chronic bronchitis, and asthma and exposure to air pollution. Interpretation requires caution regarding potential association of duration of disease with exposure status *(survival bias)*.

Survival bias may be minimized if information can be obtained on exposures that clearly preceded the first symptoms of a chronic disease such as arthritis, diabetes, or chronic bronchitis. This depends on access to medical records before the onset of a chronic disease. In addition, it may be necessary to have historical records on an individual's exposure status prior to these first medical visits, e.g. where the person lived or where the person was employed.

Terminology

Antecedent-consequent bias: occurs in cross-sectional studies when it cannot be determined if exposure preceded disease.

Prevalence: the proportion of diseased individuals in a population.

Survival bias: occurs in cross-sectional studies when the exposure influences survival time, and the distribution of that exposure will be distorted among a sample of survivors. (a.k.a. Neyman bias, incidence-prevalence bias, or selective survival bias)

Practice Questions

Answers are located at the end of this notebook

1) Which of the following health outcomes could be studied using a cross-sectional study design? Choose all that apply.

a) The prevalence of diabetes among adults in the United States in 2014
b) The prevalence of diabetes among all patients seen at a particular health clinic on one day in 2014
c) The number of new cases of diabetes diagnosed among at risk adults in the United States in 2014
d) The number of people in a population with diabetes who are obese and the number of people in a population with diabetes who are not obese, in the United States in 2014

2) Researchers are studying HIV prevalence using a cross-sectional study design. Which of the following factors may affect the researchers' assessment of HIV prevalence in their study population? Choose all that apply.

a) Changes in HIV treatment
b) Changes in HIV virulence (HIV virulence refers to the ability of the virus to cause disease)
c) Population changes (e.g. migration) due to HIV infection, for example people who leave their community to live closer to physicians specialized in treating HIV
d) Other factors which may affect survival of HIV-infected persons (e.g. changes in healthcare practices)
e) Changes in HIV diagnostics
f) Changes in HIV awareness and education

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Resources:

Delgado-Rodriguez M and Llorca J. J Epidemiol Community Health. 2004;58:635–641.

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Answers to Practice Questions

1. Answer choices a, b and d are correct. A cross-sectional study design could be used to assess the prevalence of diabetes among adults in the United States in 2014. A cross-sectional study design could also be used to assess the prevalence of diabetes among all patients seen at a particular health clinic on one day in 2014 by simply recording which patients are known to have been diagnosed with diabetes divided by the number of total patients seen that day. However, the number of new cases of diabetes diagnosed among at risk adults in the United

States in 2014 could not be assessed using a cross-sectional study. The incidence of new cases cannot be measured using a cross-sectional "snapshot" in time design. Instead, a cohort study could be used to follow at risk participants over time to see which people were newly diagnosed with diabetes in 2014. The number of people in a population with diabetes who are obese and the number of people in a population with diabetes who are not obese could be assessed using a cross-sectional design, this would be an example of an analytical cross-sectional study.

2. Answer choices b, c, d, e and f are correct. Changes in HIV virulence could affect HIV-related mortality (survival time) and, thus, would affect prevalence. Any other factors which affect survival of HIV-infected persons would also affect prevalence. Migration could also affect assessment of HIV prevalence if people migrate in or out of the study area based on their HIV status. Changes in HIV diagnostics could affect the timing of HIV diagnosis and/or how likely people are to get tested and, thus, could affect prevalence. Answer choice a is not correct because HIV is not a curable disease. HIV treatment may help an HIV-infected person live longer but will not affect whether or not they have HIV. Therefore, HIV prevalence is not affected by HIV treatment. Note that if a disease was curable with treatment then that disease's prevalence could be affected by treatment.

ERIC Notebook

Second Edition

Ecologic Studies

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Ecologic studies are studies in which the unit of observation is a group, not separate individuals, for one or more study variables. For example, exposure and risk factors are known only at the group level, such as the average air pollution concentration in different cities. The occurrence of the health outcome may also be only known at the group level, such as overall mortality rates from chronic lung disease in the same cities with measured levels of air pollution. Ecologic studies may be used to generate hypotheses of an association between exposure and a health outcome, but these studies cannot confirm causation. This is because, for example, we do not know whether those individuals who died in a particular city under observation had a higher exposure than individuals who remained alive.

Ecologic fallacy

When dealing with group level information, it is important to be aware of what is called the ecologic fallacy. This fallacy results from concluding that because an association exists between exposure and a health outcome at the group level, it therefore exists at the individual level. The cause of this fallacy is that we do not know the link between exposure and the health outcome among individuals within each group. For example, we don't know the number of diseased persons who were exposed or not exposed in the high exposure group or in the low exposure group.

Description of the ecologic fallacy

In ecologic studies, only information on aggregate measures, such as the average exposure in City A and the death rate in City A can be known. At the individual level, however, we can, for example, determine the proportion of people who died within each of the categories of exposure (low or high).

Suppose air pollution is higher in Baltimore than in Tampa, but mortality from lung disease is lower in Baltimore than in Tampa. It would be fallacious to conclude that air pollution protects against lung disease deaths. It is possible that persons dying of lung disease in Tampa may have moved from cities with high air pollution or that another risk factor for lung disease – such as smoking – is more prevalent in Tampa than Baltimore. We do not know the cumulative exposures of cases and non-cases in either city. The heterogeneity of lifetime air pollution

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

exposure among individuals in each city makes the average exposure unrepresentative of the distribution of exposure among individuals in the population.

Using the previous example, in-migrants to Tampa had previously experienced higher levels of exposure to air pollution, thus causing the aggregate level of lung disease deaths to appear higher in Tampa. The aggregate level of air pollution in Tampa does not allow us to see that there are the varying levels of exposure between lifelong residents and in-migrants, this shows how the ecologic fallacy can occur.

Examples of questions investigated by ecologic studies include:

- Is the ranking of cities by air pollution levels associated with the ranking of cities by mortality from cardiovascular disease, adjusting for differences in average age, percent of the population below poverty level, and occupational structure?
- Have new car safety measures such as passenger air bags made a difference in motor vehicle fatality rates in areas with different laws over the same period of time?
- Has introduction of newer cars with safety features such as multiple airbags made a difference in motor vehicle fatality rates in areas with a different distribution of newer versus older cars, over the same period of time?
- Are daily variations in mortality in Boston related to daily variations in particle air pollution, adjusting for season of year and temperature?
- What are the long-term time trends (1950-2012) for mortality from the major cancers in the U.S., Canada, and Mexico?

Advantages of ecologic studies

Aggregate data on exposure and health outcomes are often publicly available in state and national databases, such as the US census or from the Center for Disease Control and Prevention. Agencies of the state and federal government collect considerable data reported at the aggregate level on the economy, the environment, and the health and wellbeing of the population. Data are regularly obtained on air quality, water quality, weather conditions, the size of the population, the status of the economy, and the health of the population through surveys such as the National Health Interview Survey, National Health and Nutrition Examination, and Behavioral Risk Factor Survey. Data on the vital status of the population are obtained via birth and death registries and cancer and birth defects registries. These publicly available records provide databases for linking health outcomes with characteristics of the population, the environment, and the economy at the aggregate level.

Aggregate level data can conveniently be obtained by researchers at a low cost and can be useful for evaluating the impact of community-level interventions. Examples of interventions that may be evaluated through ecologic study designs include fluoridation of water, seat belt laws, and mass media health campaigns. Aggregate level information can be compared before and after the intervention to determine the effects of an intervention at the community level.

In addition, minimal within-community differences between exposures may exist; however exposures may differ substantially between communities, cities, states, and countries. Examples of small within-community exposure differences but large between-community differences include:

- Quality of drinking water
- Concentration of certain air pollutants such as ozone and fine particles
- Average fat content of diet (larger differences between countries than between individuals within the same city)
- Cumulative exposure to sunlight (larger differences by latitude [north-south] of residence than among individuals at the same latitude)

Ecologic studies are also useful for studying the effect of short-term variations in exposure within the same community, such as the effect of temperature on mortality.

Types of ecologic study designs

There are three main types of ecologic study designs: cross-sectional ecologic studies, time-trend ecologic studies, and solely descriptive ecologic studies.

Cross-sectional ecologic studies compare aggregate exposures and outcomes over the same time period. An example of this study design is an investigation comparing bladder cancer mortality rates in cities with surface drinking water sources that contain chlorine by-products compared to rates in cities with ground drinking water sources that contain little or no chlorine by-products.

Time-trend ecologic studies compare variations in aggregate exposures and outcomes over time within the same community. A study investigating whether hospital admissions for cardiac disease in Los Angeles increase on days when carbon monoxide levels are higher would be an example of this type of study.

Solely descriptive ecologic studies investigate disease or risk factor differences between communities at the same time, or within the same community over time. This type of study design would be used to investigate the following questions: What are the differences in lung cancer mortality among cities in North Carolina? What is the secular trend of lung cancer mortality between 1960 and 2010 for the entire state of North Carolina?

Ecologic Study Designs at a Glance

Study type	Design	Time frame
Cross-sectional	Across communities	Same time period
Time-trend	Within the same community	Over time
Descriptive	Across communities or Within the same community	At a point in time or Over time

Limitations of ecologic studies

Ecologic studies are subject to numerous biases and limitations. Most notably, these study designs are subject to the ecologic fallacy, which occurs by inferring that associations at the aggregate level are true at the individual level. Ecologic studies are also more often subject to confounding bias than are individual risk studies. *Confounding* is a mixing of the effects of other risk factors with the exposure of interest. Confounding bias may occur in an ecologic study if the confounding factor is correlated with the background rate of disease (the disease rate among unexposed persons in each study community). *Cross-level bias* occurs when the confounding factor is associated with the background rate of disease differentially across groups. The ecologic fallacy may occur as a result of cross-level bias.

Example

Suppose the association between average fat consumption and breast cancer risk is examined across communities in the US. Certain communities may have a larger percentage of women with a genetic predisposition to breast cancer than other communities. Suppose these same communities containing large percentages of women with a genetic predisposition to breast cancer are also communities with a high per capita dietary fat consumption. The results of the study will show a strong correlation between average dietary fat consumption and breast cancer mortality. The association will be inflated due to the confounding factor, genetic predisposition to breast cancer. This bias occurred because the background rate of breast cancer incidence

Time-trend ecologic studies are further limited in that an investigator cannot be confident that exposure preceded the outcome. Migration into and out of communities can also bias the interpretation of ecologic results.

Practice Questions

Answers are at the end of this notebook

1) Researchers study the community of one town in North Carolina over a 10 year period. They conduct an ecologic study and collect data on the rate of automobile accidents each year among teenagers and the percentage of teens in the town who report using their phones for texting each year. Based on their data, the researchers conclude that teenagers who report texting on their phones are more likely to be in a car accident. Which of the following are true about the researchers' conclusion? Choose all that apply.

a) The researchers' conclusion is valid

b) The researchers have incorrectly used group-level data to draw conclusions about individual teenagers

c) The researchers do not know if the teens involved in car accidents are the same teens who have reported texting on their phones, therefore, their conclusion is not valid

d) The researchers' conclusion does not account for what other factors, if any, are related to texting and related to being in a car accident

2) Reported cases of the flu are higher in city A than city B. Vaccination rates for the flu are lower in city A than city B. Which of the following are reasons why would it be incorrect to simply assume that higher vaccination in city B is what is causing city B to have fewer reported cases of the flu? Choose all that apply.

a) City A and city B may have different strains of the flu

b) City A and city B may have different proportions of people in their populations who are especially vulnerable to the flu (e.g. the elderly, children and pregnant women)

c) City A and city B may have differences in health care accessibility, leading to differences in testing for the flu and diagnosis of the flu

d) City A and city B may have different climates, leading to differences in how/where people come into contact with each other. This may affect flu transmission rates

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. Answer choices b, c and d are correct. In the example, the researchers have used overall data on car accidents for a town and overall data on the prevalence of teens who reported texting to draw a conclusion about individuals. The researchers actually do not have any data on whether the teens involved in car accidents were the same teens who reported texting. In addition, there may be other factors involved that explain the study results such as another confounding variable that is related to being in a car accident and related to reporting texting.
2. All answer choices are correct. City B may have a different strain of the flu circulating, which could affect transmission rates, morbidity/mortality and whether or not an infected person seeks medical care. There may be more people in city A who are especially vulnerable to the flu and these vulnerable groups may experience greater illness, leading to more visits to a health provider and more diagnoses. If health care accessibility varies between the 2 cities and/or knowledge of testing and diagnosis for the flu differs between the 2 cities, this may affect whether infected persons seek health care and may affect which tests they receive. Such variability may lead to the appearance of greater or fewer cases. Finally, the climate may influence social mixing patterns leading to greater virus transmission in the city where people spend more time together indoors.

ERIC Notebook

Second Edition

Incident vs. Prevalent Cases and Measures of Occurrence

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

To determine which factors impact health outcomes at the population level, epidemiologists employ a number of different study designs. These designs are used to examine the relationship between exposures (or determinants) and health outcomes. A health outcome may be a disease, condition, death, event or a change in health status or behavior. For example, in addition to diseases, we may study health events such as injuries or the occurrence of an "event" such as preterm birth. Persons who experience the outcome of interest are commonly referred to as cases.

One of the first things to consider when developing a study is whether you will measure prevalent or incident cases.

Prevalent cases are all individuals living with the outcome of interest within a specified timeframe, regardless of when that person was diagnosed or developed the health outcome.

Example

In a study of prevalent cases of diabetes with a one year time period, anyone who has diabetes during the one year study period would be counted as a case.

These prevalent cases would include both people who have diabetes at the outset of the study year as well as any who developed diabetes over the course of the study.

Incident cases are all individuals who change in status from non-disease to disease – or from one state of a health outcome to another – over a specific period of time. In other words, "incidence" refers to the occurrence of new cases.

Example

For example, in a study of incident cases of diabetes with a one year time period, only those who developed diabetes over the course of the one year study period are considered incident cases.

Measures of frequency

Prevalence

Prevalence is the proportion of a population living with a specific health outcome within a specified time. It is the only measure of occurrence calculated with prevalent cases. To calculate prevalence, the number of prevalent cases (numerator) is divided by the total population at risk

(denominator.) The total population at risk denominator includes the prevalent cases. Prevalence is often reported as a percentage.

Prevalence = Prevalent cases / Total population

Depending on the type of prevalence being calculated, the denominator can be either an average of the population over time or a single measurement at a specific point of time.

Prevalence can either be calculated as a *point prevalence* or *period prevalence.* A point prevalence is calculated with data from one specific point in time, while a period prevalence is calculated over a range of time.

Prevalence is directly affected by the incidence and duration of the health outcome under study, which makes it a poor choice for diseases or outcomes with a short duration or high mortality rate.

Example

For example, *Vibrio vulnificus* – a disease caused by consumption of raw shellfish – has a low incidence and short duration. Therefore, the few new (incident) cases that arise will remain prevalent in the population for only a short time before the cases recover or die. However, for a disease like diabetes, which has a higher risk or rate and longer duration, the prevalence will be higher than the risk or rate and is a valuable measure of the burden of disease in the population.

Prevalence = Rate x Duration

Risk

Like prevalence, risk is also a measure of the extent of a health outcome in a population. However, unlike prevalence, risk is the proportion of an at-risk population that develops a specific health outcome within a specified amount of time. The numerator for risk is incident cases, and the denominator includes only those at-risk of developing the outcome of interest at the beginning of study follow-up.

Example

If the disease under study is ovarian cancer, which obviously only affects women, the denominator should consist only of women in the population who, at the start of study observation, do not have ovarian cancer and are capable of developing ovarian cancer.

Risk = Incident cases / Population at-risk

Risks are often reported as a scaled value, such as cases per 1,000; 10,000; or 100,000 population.

Rate

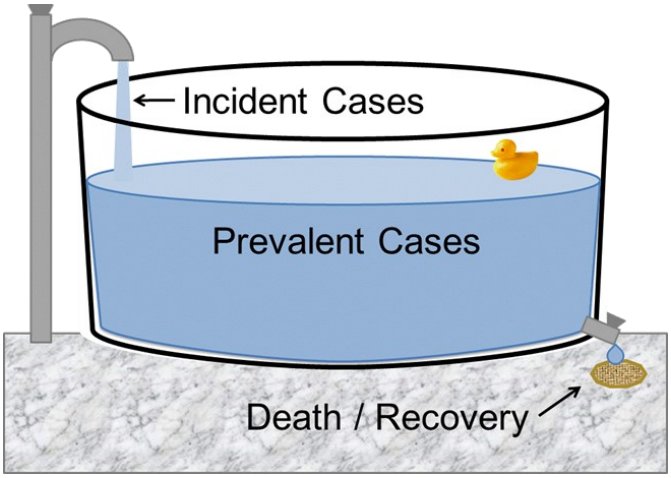
Rate is another measure of health outcome occurrence calculated with incident cases of the health outcome. However, the denominator for a rate is the total amount of person-time at-risk. Person-time is an estimate of the actual time-at-risk – in years, months, or days – that all participants contributed to a study. In its simplest form, person-time is a sum of each study participant's time at risk before experiencing the outcome of interest or exiting the study. This is a better estimate of the true at-risk population because it excludes time for participants who are no longer eligible to experience the outcome of interest. Thus, rates are a better reflection of health outcome occurrence in a dynamic population, where participants may exit the study or become no longer at-risk.

Rate = Incident cases / Total person-time at-risk

The unit for a rate is "cases per person-time." Rates are often reported as a scaled value with a time unit relevant for the study, such as cases per 1,000 person-years, 12,000 person-months, or 365,000 person-days.

Rates are favored if the rapidity with which new cases of the health outcome or new events are occurring in the population is of interest.

Comparing Measures of Occurrence



The image of the bathtub below graphically represents the relationships between prevalence, risk, and rate.

In this analogy, prevalence is the proportion of the tub (the total population) filled with any water (prevalent cases and incident cases). Risk is the proportion of the tub filled with new, flowing water (incident cases). Rate is a measure of how quickly the water flows into the tub. Prevalent cases only leave the prevalence pool by either recovery, death, migration out of the population or loss of study follow-up via the bathtub drain.

Practice Questions

Answers are located at the end of this notebook.

1) If researchers were studying the risk of a woman having a baby born preterm in the United States in 2013, what would the at-risk population be? Assume for this question that a pre-term birth is any birth before 39 weeks gestation. Choose the one best answer.

a) All women in the United States in 2013
b) All pregnant women in the United States in 2013
c) The actual number of babies born preterm in the United States in 2013
d) All babies born in the United States in 2013
e) All pregnant women in the United States whose due dates would mean that the baby could potentially be born pre-term in the year 2013

2) Now researchers want to study the risk of head injuries that occur while a person was riding a bicycle, in the United States between 2000-2013. How would the risk denominator ideally be calculated? Choose the one best answer.

a) All people living in the United States in the time period 2000-2013
b) All people who rode a bicycle in the United States in the time period 2000-2013
c) All people who owned or had access to a bicycle in the United States in the time period 2000-2013
d) The actual number of documented head injuries due to bicycle accidents in the United States in the time period 2000-2013
e) All people who had a head injury in the United States in the time period 2000-2013

Terminology

Prevalent cases – all individuals living with the health outcome of interest within a specified timeframe, regardless of when that person was diagnosed or developed the health outcome

Incident cases – all individuals who change in status from one state of health to another (such as non-disease to disease) over a specific period of time

Prevalence – the proportion of a population living with a specific health outcome within a specified timeframe

Risk – the proportion of an at-risk population that develops a specific health outcome within a specified amount of time

Rate – the frequency of incident cases per unit of person-time

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. The best answer was e: All pregnant women in the United States whose pregnancy due date would mean that the baby could potentially be born pre-term in the year 2013. Since the researchers are studying risk, they need to measure incident cases. The denominator must include only those at-risk of having the outcome of interest during the defined time-period of interest (the year 2013).

Answer choice a (All women in the United States in 2013) is not the best choice because only pregnant women—and not all women in general—are at-risk of having a preterm baby born. Answer choice b (All pregnant women in the United States in 2013) is a good choice but is not the best answer choice among those listed. This is because a women who is pregnant in 2012 may be at-risk of giving birth to a pre-term baby is the calendar year 2013. So the researchers would not want to limit their at-risk population just to women who are pregnant in the United States in 2013. Answer choice c (The actual number of babies born preterm in the United States in 2013) is incorrect because this is not the at-risk population needed for a risk measure since this is a count of babies already born preterm. Similarly, answer choice d (All babies born in the United States in 2013) is incorrect because it does not represent the at-risk population that would need to be followed over time.

2. The best answer was b: All people who rode a bicycle in the United States in 2013. In your risk denominator, you want to only include people actually at-risk of the outcome. In order to have a head injury that arises from riding a bicycle, a person would actually have to be riding a bicycle, otherwise they would never be at-risk for the outcome under study.

Answer a (All people living in the United States in the time period 2000-2013) is not the best choice because all of these people did not ride a bicycle. Answer c (All people who owned or had access to a bicycle in the United States in the time period 2000-2013) is incorrect for the same reason. Just because someone has access to a bicycle does not mean they actually rode one. Answer d (The actual number of documented head injuries due to bicycle accidents in the United States in the time period 2000-2013) is not correct because the researchers are studying incident cases among all people at risk for the outcome. The group at-risk is all people who rode a bicycle and not just those who actually had a head injury. Answer e (All people who had a head injury in the United States in the time period 2000-2013) is incorrect because the researchers were not studying head injuries in general, they were studying head injuries arising from riding a bicycle specifically.

ERIC Notebook

Second Edition

Sources of Systematic Error or Bias: Information Bias

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Information bias is one type of systematic error that can occur in epidemiologic studies. *Bias* is any systematic error in an epidemiologic study that results in an incorrect estimate of the association between exposure and the health outcome. Bias occurs when an estimated association (risk ratio, rate ratio, odds ratio, difference in means, etc.) deviates from the true measure of association.

Bias is caused by systematic variation, while chance is caused by random variation. The consequence of bias is systematic error in the risk ratio, rate ratio, or odds ratio estimate. Bias may be introduced at the design or analysis phase of a study. We should try to eliminate or minimize bias through study design and conduct.

Major types of systematic error include the following:

- Selection bias
- Confounding bias
- Information bias

In this issue we present information bias. Selection bias and confounding are covered in separate ERIC Notebooks.

Information bias is a distortion in the measure of association caused by a lack of accurate measurements of key study variables. Information bias, also called measurement bias, arises when key study variables (exposure, health outcome, or confounders) are inaccurately measured or classified. Bias in the risk ratio, rate ratio, or odds ratio can be produced even if measured errors are equal between exposed and unexposed or between study participants that have or do not have the health outcome.

Non-differential misclassification

Non-differential misclassification occurs if there is equal misclassification of exposure between subjects that have or do not have the health outcome or if there is equal misclassification of the health outcome between exposed and unexposed subjects. If exposure or the health outcome is dichotomous, then non-differential misclassification causes a bias of the risk ratio, rate ratio, or odds ratio towards the null.

Non-differential misclassification of exposure status

Non-differential misclassification of exposure status in a case-control study occurs when exposure status is

equally misclassified among cases and controls. Non-differential misclassification in a cohort study occurs when exposure status is equally misclassified among persons who develop and persons who do not develop the health outcome.

Non-differential misclassification of health outcome status occurs in a case-control study when the health outcome status is equally misclassified among exposed and unexposed subjects. Non-differential misclassification of the health outcome status occurs in a cohort study when a study subject who develops the health outcome is equally misclassified among exposed and unexposed cohorts.

Effect of non-differential misclassification of exposure

Non-differential misclassification biases the risk ratio, rate ratio, or odds ratio towards the null if the exposure classification is dichotomous, i.e., either exposed or unexposed. If exposure is classified into 3 or more categories, intermediate exposure groups may be biased away from the null, but the overall exposure-response trend will usually be biased towards the null.

Effect of non-differential misclassification of the health outcome

In most cases, non-differential misclassification of the health outcome will produce bias toward the null, i.e. the risk ratio, rate ratio or odds ratio will be biased towards 1.0. If errors in detecting the presence of the health outcome are equal between exposed and unexposed subjects (i.e. sensitivity is less than 100%) but no errors are made in the classification of health outcome status (i.e. specificity is 100%), the risk ratio or rate ratio in a cohort study will not be biased, but the risk difference will be biased towards the null.

Effect of non-differential misclassification of health outcome status

If no errors are made in detecting the presence of the health outcome (i.e. 100% sensitivity), but equal errors are made among exposed and unexposed in the classification of health outcome status (i.e. specificity less than 100%), the risk ratio, rate ratio, and risk difference (as applicable) will be biased towards the null.

Combined errors in both sensitivity and specificity further increase the bias towards the null, but specificity errors produce larger biases overall.

Differential misclassification

Differential misclassification occurs when misclassification of exposure is not equal between subjects that have or do not have the health outcome, or when misclassification of the health outcome is not equal between exposed and unexposed subjects.

Differential misclassification causes a bias in the risk ratio, rate ratio, or odds ratio either towards or away from the null, depending on the proportions of subjects misclassified.

Effect of differential misclassification of exposure or health outcome

Differential misclassification of the exposure or health outcome can bias the risk ratio, rate ratio, or odds ratio either towards or away from the null. The direction of bias is towards the null if fewer cases are considered to be exposed or if fewer exposed are considered to have the health outcome. The direction of bias is away from the null if more cases are considered to be exposed or if more exposed are considered to have the health outcome.

The effect of differential misclassification of the exposure or health outcome can bias the risk ratio, rate ratio, or odds ratio in either direction. The direction of bias is towards null if fewer cases are considered to be exposed or if fewer exposed subjects are considered to have the health outcome. The direction of bias is away from the null if more cases are considered to be exposed or if more exposed subjects are considered to have the health outcome.

Interviewer bias

Interviewer bias is a form of information bias due to:

1. lack of equal probing for exposure history between cases and controls (exposure suspicion bias); or

2. lack of equal measurement of health outcome status between exposed and unexposed (diagnostic suspicion bias)

Solutions:

1. blind data collectors regarding exposure or health outcome status
2. develop well standardized data collection protocols
3. train interviewers to obtain data in a standardized manner
4. seek same information about exposure from two different sources, e.g. index subject and spouse in case-control study

Recall or reporting bias

Recall or reporting bias is another form of information bias due to differences in accuracy of recall between cases and non-cases or of differential reporting of a health outcome between exposed and unexposed.

Cases may have greater incentive, due to their health concerns, to recall past exposures. Exposed persons in a cohort study may be concerned about their exposure and may over-report or more accurately report the occurrence of symptoms or the health outcome.

Solutions:

1. add a case group unlikely to be related to exposure
2. add measures of symptoms or health outcomes unlikely to be related to exposure

Complications in predicting direction of misclassification bias

Misclassification of confounders results in unpredictable direction of bias. Non-differential misclassification of a polychotomous exposure variable (3 or more categories) may result in bias away from null, though this is less likely than bias towards the null.

Non-differential misclassification of a health outcome limited to a loss of sensitivity of detecting the health outcome without any loss in specificity does not bias toward null, whereas a loss of specificity always biases toward the null.

Conclusions

Some inaccuracies of measurement of exposure and health outcome occur in all studies.

If a positive exposure-health outcome association is found and non-differential measurement errors are more likely than differential ones, measurement error itself cannot account for the positive finding since non-differential error nearly always biases towards the null.

Strive to reduce errors in measurement:

1. develop well standardized protocols
2. train interviewers and technicians well
3. perform pilot studies to identify problems with questionnaires and measuring instruments
4. attempt to assess the direction of bias by considering likelihood of non-differential or differential misclassification

Terminology

Information bias: A distortion in the measure of association caused by a lack of accurate measurements of exposure or health outcome status which can result from poor interviewing techniques or differing levels of recall by participants.

Non-differential misclassification: Equal misclassification of exposure between subjects that have or do not have the health outcome, or equal misclassification of the health outcome between exposed and unexposed subjects.

Differential misclassification: Unequal misclassification of exposure between subjects that have or do not have the health outcome, or unequal misclassification of the health outcome between exposed and unexposed subjects.

Practice Questions

Answers are at the end of this notebook

1) Researchers conduct a case-control study. The following table shows the true classification of exposure and the health outcome (Note: these data are hypothetical and typically researchers would not know the true unbiased distribution of exposure and outcome).

	Have health out-come	Do not have health out-come
Exposed	200	210
Unex-posed	340	500

a) Calculate the odds ratio

Now imagine that 50 people with the health outcome were misclassified as being unexposed and 20 people with the health outcome were misclassified as being exposed.

b) Create the corrected 2x2 table

	Have health out-come	Do not have health outcome
Exposed		
Unex-posed		

c) Calculate the odds ratio for the corrected table
d) In which direction was the misclassification bias?

2) Researchers conduct a case-control study of the association between the diet of young children and diagnosis of childhood cancer, by age 5 years. The researchers are worried about the potential for recall bias since parents are being asked to recall what their children generally ate, over a period of 5 years. Which of the following potential control groups would be most likely to reduce the likelihood of recall bias?

a) Parents of children with no known health problems
b) Parents of children with other known, diagnosed serious health problems (aside from childhood cancer)
c) Parents of children with other known, diagnosed minor health problems

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

Dr. Steve Marshall and Dr. Jim Thomas Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, The University of North Carolina at Chapel Hill, Department of Epidemiology

Dr. David Richardson Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013, The

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of

Answers to Practice Questions

1.

a) Calculate the odds ratio

Odds ratio = (200*500) / (340*210) =1.4

b) Create the corrected 2x2 table

	Have health out-come	Do not have health outcome
Exposed	230	210
Unex-posed	310	500

c) Calculate the odds ratio for the corrected table

Odds ratio = (230*500) / (310*210) =1.8

d) In which direction was the misclassification bias?

The bias was away from the null (the null value is 1.0).

2. Answer choice b is the best choice. Researchers should aim to have similar recall bias between both the case and control groups. Parents of children who have childhood cancer, which is a serious health problem, are likely to be quite concerned about what may have contributed to the cancer. Thus, if asked by researchers, these parents are likely to think very hard about what their child ate or did not eat in their first years of life. Parents of children with other serious health problems (aside from cancer) are also likely to be quite concerned about any exposure that researchers ask about. Therefore, these parents can be expected to recall exposures in a way that is more comparable with parents of children who have cancer. In contrast, parents of children who have no health problems or parents of children with only minor health problems are less likely to be concerned with carefully recalling any exposures.

ERIC Notebook

Second Edition

Randomized Controlled Trials (Experimental Studies)

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

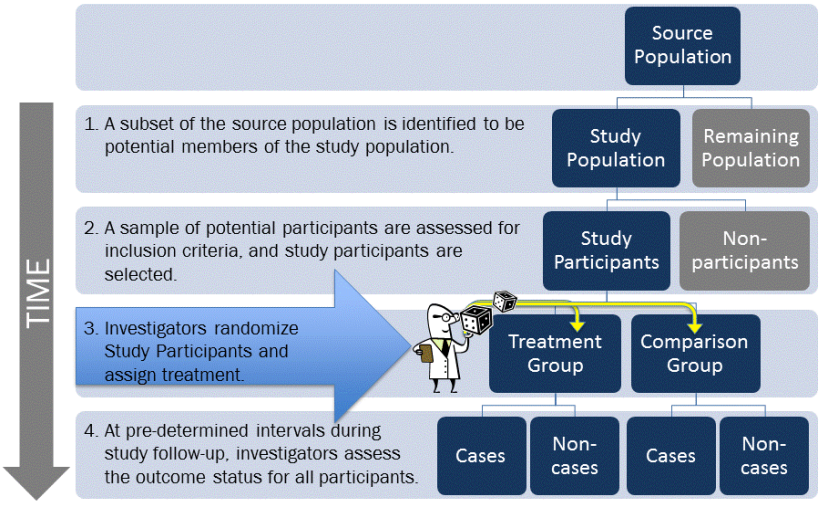
What are randomized trials?

Randomized trials are epidemiological studies in which a direct comparison is made between two or more treatment groups, one of which serves as a control for the other. Study subjects are randomly allocated into the differing treatment groups, and all groups are followed over time to observe the effect of the different treatments. The control group may either be untreated (*placebo-controlled*) or undergo a "gold standard" established regimen against which the new regimen will be assessed (*active-controlled*). Randomized trials provide the most direct evidence for causality. However, they are also fraught with a number of additional considerations not present for observational research.

For example, unless researchers are genuinely uncertain about the potential harms or benefits of a treatment, it is unethical to assign it to one group of people while withholding it from others (*equipoise*). This limits the types of questions that can be answered using experimental studies.

A *placebo-controlled* randomized trial might compare the effect of vitamin E supplement in one group of schizophrenia patients (the treatment group) against the effects of a placebo on a separate group of schizophrenia patients (the control group).

An *active-controlled* randomized trial might compare diabetic patients with implanted insulin pumps against diabetic patients who receive multiple insulin injections (the control group).



Randomized Trial Study Design

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Randomization

Randomization avoids bias by eliminating baseline differences in risk between treatment and control groups. Randomization, if done properly, should make both groups similar in terms of the distribution of risk factors, regardless of whether these risk factors are known or unknown (thus eliminating confounding due to both measured and unmeasured variables). The larger the randomized groups, the greater the probability of equal baseline risks. However, participants in RCTs are often not representative of the target population, which introduces selection bias and limits generalizability.

Methods of randomization

There are different ways to randomize study participants into treatment groups. A simple way to randomize would be to roll a die or use a random number table to allocate individuals into the different groups. Another way investigators randomize study participants is through stratified random allocation. Under this method, the investigator first stratifies the participants by a baseline risk factor (i.e., smoking status) then randomizes the subjects in each stratum into either the treatment or control group. Stratified random allocation is appropriate when the investigator wants to be sure that a strong external risk factor is equalized at baseline between treatment and control groups.

Types of randomized trials

The two general types of randomized trials are *clinical trials* and *community trials*, with randomized clinical trials being by far the more common. A randomized clinical trial is an experiment with patients as subjects. The goal is to find an effective treatment for a disease or to evaluate an intervention to prevent the progression of a disease. Randomized clinical trials are often used to evaluate the efficacy of new drugs against standard treatments or against placebos, but they are also used to evaluate other therapeutic procedures such as a new form of surgery, a dietary regimen, or an exercise program for persons with pre-existing disease. Most often, patients who already have some specific disease are the subjects of study in clinical trials. However, at times, subjects who are at high risk for a specific disease are entered into a randomized clinical trial to assess the efficacy of a drug to prevent the disease. For example, women with a family history of breast cancer may be entered into a clinical trial to study the effect of the drug tamoxifen on the prevention of breast cancer.

A community trial is also an experiment, but differs from clinical trials in that an entire community, rather than an individual patient, is the unit of observation. For example, water fluoridation was evaluated by experimentally assigning entire communities to have their public water supply fluoridated or not fluoridated. Units of observation for a community trial may be a town or city, a factory or office, a classroom or an entire school. All persons in the same unit of observation are experimentally exposed to the same intervention although it is not certain that all persons in the unit will be equally exposed, e.g. that they will drink the fluoridated water coming from their taps. Several community trials have been conducted to evaluate the effectiveness of mass media campaigns to prevent heart disease by encouraging more exercise, less use of tobacco products, and other lifestyle modifications.

Blinding or masking

Sometimes in clinical trials, participants, statisticians, and even investigators, are made unaware of whether the participants are part of the treatment or control group. When only study participants are unaware of their treatment status, but investigators and analysts are aware of treatment status, the trial is called *single-blinded*. When both the participants and the investigators are blinded as to the treatment status of the participants the trial is termed *double-blinded*. A *triple-blinded* trial is when subjects, investigators, and independent statisticians are kept unaware of subject treatment status.

Blinding the study participants by using placebos, or a sham treatment, is common practice in clinical trials. The

placebo effect occurs when participants report a favorable response when no treatment, but only placebo, is administered. Another bias that is prevented by blinding of subjects is *post-randomization confounding bias* where subjects' awareness of intervention may motivate them to be more cooperative or otherwise change their behavior. This motivation may correlate with other risk factors for the intended effect, thus destroying the design advantage of randomization.

Example

If individuals participating in a clinical trial to study the efficacy of a new weight loss drug are aware that they are receiving the weight loss drug, they may more closely comply with the prescribed study diet.

Another bias that is controlled for by blinding the subjects as to their treatment status is *selection bias*, or group differences in loss to follow-up. Symptoms of disease or side effects of the treatment may influence rates of loss to follow-up in subjects aware of their treatment status.

Bias due to differences in reporting of symptoms, a type of *information bias,* is also controlled by a double-blinded study. Study subjects who are aware of their treatment status may differentially report symptoms or side effects. Likewise, staff or statisticians may differentially evaluate subjects if they are aware of treatment status.

Example

In a study of the effects of a new drug on severity of migraines in which study members know their treatment status, the treated study members may believe that the drug will work and, therefore, report less severe migraines. If the investigator in this study knows the treatment status of the subjects, then that investigator may scrutinize the severity of the migraines in treated subjects more than that of the untreated subjects.

Additional threats to the validity of a randomized trial

Limiting the analysis to compliant subjects can create bias if compliance is correlated with other risk factors for the treatment effect. Analyzing the results without regard to subject compliance (called "intention-to-treat" analysis) can help to avoid this bias. That is, subjects should be included in the analysis whether or not they adhered to their treatment (or control) regimen.

Example

Suppose that in a clinical trial to look at the relationship between diet and risk of cancer, subjects were randomized to either a cancer-prevention diet or to a placebo diet. Suppose again that in the treatment group, those subjects with gastrointestinal symptoms that are precursors of cancer, were less compliant with their diet than subjects without symptoms. Exclusion from the analysis of the less compliant subjects would bias the results towards reporting a greater effect of the cancer-prevention diet. Only those subjects who were not at risk or who were at low risk of cancer would be included in the analysis. The appropriate analysis should include all persons originally assigned to their treatment group, whether or not they adhered to the treatments.

When noncompliant subjects are selectively excluded from an analysis, the benefit of randomization is lost, because unmeasured confounding factors may be associated with the lack of compliance.

Treatment crossover

Crossover, either planned or unplanned can create biases in experiments. In a *planned crossover*, group A (subjects treated with the new drug) and group B (subjects treated with a standard drug) would be switched to the other treatment at the midpoint of the trial. Two of the problems experienced with this experimental design are *carryover effects* and *diminished interest*. Carryover effects occur when the effects of the first drug last into the second half

of the study when the subjects are receiving the other treatment. Bias may also occur if there is diminished interest or lack of compliance in the second half of the study.

Unplanned crossovers occur when a clinician decides to switch a study member from the control to the treatment group, or vice versa, e.g. surgery vs. medical treatment for coronary artery disease. An unplanned crossover negates the benefit of randomization and introduces bias if switching is related to risk of the outcome.

Loss to follow-up

Neither randomization nor blinding can prevent *differential loss to follow-up*, or more subjects dropping out in one treatment group than in another. Bias is introduced if the rate of loss to follow-up is correlated with both exposure to the treatment and exposure to other risk factors for the outcome.

Threats to validity:

Loss to follow-up
Non-compliance
Crossovers

Analysis strategies to avoid bias

For purposes of analysis, study subjects should be kept in the original randomized group, even if they were lost to follow-up, switched to the other treatment group, or were non-compliant (the "intention-to-treat" principle). Analysis of any non-random subgroups threatens the validity of the study.

Practice Questions

Answers are located at end of this notebook.

1) Researchers conducted a multi-year ongoing randomized controlled trial of the association between daily meditation (such as relaxation techniques) and health behavior among patients following a skin cancer diagnosis. Researchers randomly allocated study participants into 2 groups. The first patient group received weekly classes on meditation practices as well as a self-taught manual on meditation. The second patient group received only the self-taught manual on meditation. Data were collected at set intervals following the intervention to assess the patients' health behaviors. Studied health behaviors included data on the patient's diet, exercise, and mental health. Note: this is a hypothetical example.

a) In this example, why was it ethical for the researchers to allocate one group to receive weekly classes on meditation practices as well as a self-taught manual on meditation while the other group received only the self-taught manual on meditation?

b) Which of the following may bias the analysis? Choose all that apply.

a) Changes over time in how the health behaviors were defined and assessed

b) Inability to blind the researchers regarding which of the 2 meditation interventions each patient received

c) Lack of use of a separate untreated control group (e.g. a group that received no meditation intervention at all)

d) Patients that are not compliant with their assigned group (e.g. patients assigned to just the self-taught manual but who really want to be as healthy as possible so they show up at the weekly classes on meditation practices)

c) In this example study, would stratified random allocation have been useful?

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Answers to Practice Questions

1) In this example, researchers must be genuinely uncertain about the potential benefits of mediation in order for it to be ethical to assign different meditation interventions to different groups. If the effect of meditation on health behavior has not been extensively studied in this study population, then that would make it ethical to conduct this study.

b) Answer choices a, b, and d are correct. If researchers make changes in how they assess health behaviors over time, that would bias the results of the study. If researchers are not able to be blinded in regard to which patients are in which intervention group, this may lead researchers to assess or question the study participants about their health behaviors in different ways depending on which group they are in. Randomization works to eliminate baseline differences in risk between the 2 groups being compared. If some patients are not compliant with the group they were randomized to, this can negate the benefits of randomization. Lack of use of a separate untreated control group does not introduce bias into the study. The comparison groups are chosen based on the researchers' study hypothesis and based on what is ethical. If the researchers wanted to study the effect of weekly meditation classes + a self-taught manual on mediation versus just the self-taught manual on meditation then this is a valid study hypothesis.

c) Stratified random allocation may have been useful. Stratified random allocation is when the researchers first stratify participants by a baseline risk factor and then randomize the subjects in each stratum into the 2 comparison groups. Stratified random allocation can be done when the researchers want to be certain that a strong external risk factor is equalized at baseline between the 2 comparison groups. In our hypothetical example, researchers may have hypothesized that those participants who had strong family support would be more likely to learn and adopt meditation practices and more likely to have positive health behaviors. So the researchers could have first stratified participants based on the level of family support they reported and then, after that, randomized subjects into the 2 comparison groups.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

ERIC Notebook

Second Edition

Risk and Rate Measures in Cohort Studies

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Cohort studies are longitudinal studies where an exposed and an unexposed group (or less exposed group) are followed forward in time to find the incidence of the outcome of interest (e.g. disease, death, or change in health). Two measures of incidence are risks and rates. Risks and rates can be

further manipulated to provide additional information on the effects of the exposure of interest, such as risk or rate ratios, risk or rate differences, and attributable risk fractions.

Risk is defined as the number of new cases divided by the total population-at-risk at the beginning of the follow-up period. An individual's risk of developing the outcome of interest is

$$\text{Risk} = \frac{\#\ of\ new\ cases}{total\ \#\ individuals\ at\ risk}$$

measured.

A rate is the number of new cases of a health outcome divided by the total person-time-at-risk for the population. Person-time is calculated by the sum total of time all individuals remain in the study without developing the outcome of interest (the total amount of time that the study members are at risk of developing the outcome of interest). Person-time can be measured in days, months, or years, depending on the unit of time that is relevant to the study. A rate measures the rapidity of health outcome occurrence in the population.

$$\text{Rate} = \frac{\#\ of\ new\ cases}{total\ person\text{–}time\ at\ risk}$$

Two-by-two tables are generally used to organize the data from a study as shown below.

	Disease	No disease	Total
Exposed	a	b	a+b
Unex-posed	c	d	c+d
Total	a+c	b+d	a+b+c+d

Risk ratios

When risks are computed in a study, the risk ratio is the measure that compares the $\text{Risk}_{\text{exposed}}$ to the $\text{Risk}_{\text{unexposed}}$. The risk ratio is defined as the risk in the exposed cohort (the index group) divided by the risk in the unexposed cohort (the reference group). A risk ratio may vary from zero to infinity.

$$\text{Risk Ratio} = [a / (a+b)] / [c / (c+d)]$$

$$\text{Risk Ratio} = Risk_{exposed}/Risk_{unexposed}$$

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

For example, suppose researchers conduct a cohort study and gather the following data on the effects of gasoline fume exposure on respiratory illness among automotive workers.

	Disease	No disease	Total
Exposed	60	140	200
Unex-posed	25	175	200
Total	85	315	400

In this study, the risk in the exposed group is 60/200, or 0.30 cases per person (30 cases per 100 people), and the risk in the unexposed group is 25/200, or 0.125 cases per person (13 cases per 100 people). Therefore, the risk ratio is 0.30/0.125, or 2.4. A risk ratio of 2.4 implies that the exposed group has 2.4 times the risk of developing respiratory illness as the unexposed group.

Rate ratio

When rates are computed in a study, the rate ratio is the measure that compares the $Rate_{exposed}$ to the $Rate_{unexposed}$. The rate ratio is defined as the rate of health outcome occurrence in the exposed cohort (the index group) divided by the rate of health outcome occurrence in the unexposed or less-exposed cohort (the reference group).

$$\text{Rate Ratio} = Rate_{exposed}/Rate_{unexposed}$$

A rate ratio measure also may show whether the exposure was preventive, harmful, or had no effect on the rate of health outcome in the exposed.

If in the previous example, the person-time-at- risk that each automotive worker contributed to the study had been recorded then the table might have looked like the following:

	Disease	No disease	Person-years at risk
Exposed	60	140	175
Unexposed	25	175	188
Total	85	315	363

In this study, the rate in the exposed cohort is 60/175 person- years, or 0.34 cases/person-year. The rate in the unexposed cohort is 25/188 person-years, or 0.13 cases/ person-year. The rate ratio in this study is 0.34/0.13, or 2.6, which is higher than the rate ratio calculated above. This rate ratio reveals that respiratory illness among workers exposed to gasoline fumes is developing at 2.6 times the rate that respiratory illness is developing among workers not exposed to gasoline fumes.

An exposure may be preventive (e.g., vitamin intake) or harmful (e.g., toxic chemical exposure). Confounding, which you will read about in another ERIC notebook issue, is one type of systematic error that can occur in epidemiologic studies. Confounding can cause an over- or under-estimate of the observed association between exposure and a health outcome. Assuming there are no other factors that may confound the association, a risk ratio less than 1 indicates that the risk in the exposed (index) group is less than the risk in the unexposed or less -exposed (reference) group, and therefore, the exposure is preventive. A risk ratio or rate ratio that equals 1 (the null value) indicates that there is no difference in risk or rates between exposed and unexposed groups. A risk ratio greater than one indicates that the risk in the exposed is greater than the risk in the unexposed, and, therefore, the exposure is harmful.

The following table may be applied to both risk and rate ratios.

Risk ratio or	Exposure
<1	Exposure is protective
=1	Exposure is neither preventive nor harmful (null association)
>1	Exposure is harmful

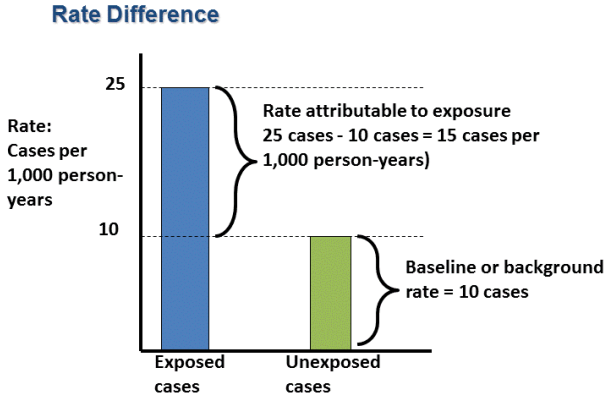
The farther away the risk ratio or rate ratio is from the null value of one, the greater the effect of exposure is on the study group. This is shown in the following diagram.

RR=risk ratio or rate

Risk and Rate differences

Difference measures are absolute measures of disease burden and helpful for public health program planning. Ratio measures are relative measures; they are commonly used to research the etiology of disease. Risk and rate differences answer the question "how much disease is attributed to the exposure?" Risk difference is defined as the risk in the exposed minus the risk in the unexposed (Equation 1). The risk difference is the *excess* risk of disease among the exposed population. Rate difference uses a similar equation (Equation 2.) In the past, risk difference was called "attributable risk"; sometimes attributable risk is still used. "Attributable fraction among the exposed" is the risk difference reported as a percent of the exposed population (Equation 3). "Attributable proportion" or "attributable risk percent" are alternative terms for Equation 3. Lastly, the attributable fraction among the total population (Equation 4) answers the question "what proportion of disease in the *total* population is associated with the exposure?" See figure, terms, and equations defined below.

Figure



Adapted from Gordis L. Epidemiology, p. 156

Equation 1: Risk Difference= $Risk_{exposed}$ - $Risk_{unexposed}$

Equation 2: Rate Difference = $Rate_{exposed}$ - $Rate_{unexposed}$

Equation 3: Attributable fraction among the exposed:

$$RD\%(exposed) = \frac{\text{risk in exposed} - \text{risk in unexposed}}{\text{risk in exposed}} \times 100$$

Equation 4: Attributable fraction among the total population:

$$RD\%(\text{total population}) = \frac{\text{risk in total pop} - \text{risk in unexposed}}{} \times 100$$

Practice Questions

Answers are at the end of this notebook

1) Researchers conduct a prospective cohort study to assess the association between dietary supplements and cognitive ability among children. A total of 500 children age 12-17 years who take an omega-3 fatty acid supplement are compared with 500 children age 12-17 years who do not take an omega-3 fatty acid supplement. Researchers follow the children for 2 years. During this time, 300 children who take the supplement earn what is classified as a "high" score on a cognitive test while 200 children who do not take the supplement earn what is classified as a "high" score on the same cognitive test.

a) Construct a 2x2 table from the information presented above

b) The risk difference is:

c) The attributable fraction among total population is:

d) The attributable fraction among the exposed is:

2) Researchers conduct a prospective cohort study of the association between working at a high-stress job and depression. A total of 5000 adults are followed for an average of 10 years. Among these 5000 adults, 2500 had a high-stress job while 2500 had a low-stress job. Participants had yearly health checks. A total of 250 incident cases of depression were diagnosed in the high-stress job group while 90 incident cases of depression were diagnosed in the low-stress job group. Assume all cases of depression were diagnosed at the end of year 5 of follow-up.

a) Construct a 2x2 table from the information above

b) Rate exposed =

Rate unexposed =

c) Rate ratio =

d) Rate difference=

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health. 1998 Jan;88(1):15-9. No abstract available. Erratum in: Am J Public Health. 2008 Dec;98(12):2119

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1.

a)

	Scored "high" on exam	Did not score "high" on exam	Total
Exposed to supplement	300	200	500
Unexposed to supplement	200	300	500
Total	500	500	1000

b) Risk exposed = 300/500= 0.6 cases/person
Risk unexposed = 200/500 = 0.4 cases/person
Using equation 1, we calculate the Risk difference as 0.6 – 0.4 = 0.2 cases/person or 20 cases per 100

c) Attributable fraction in total population (equation 4)
Risk in Total population 500/1000 = 0.5 cases/person
Attributable risk in total population= (0.5-0.4)/0.5= 0.2 cases/person or 20 cases per 100

d) Attributable fraction in exposed (equation 3)
Risk exposed = 300/500= 0.6 cases/person
Risk unexposed = 200/500 = 0.4 cases/person
Attributable risk in exposed = (0.6-0.4)/0.6=0.33 cases/person or 33 cases per 100

2.

a)

Study group	Depression	No depression	Total person-years
High-stress job	250	2250	(250*5+2250*10) = 23,750
Low-stress job	90	2410	(90*5+2410*10)= 24,550
Total	340	4660	48, 300

b) Rate exposed = 250/23,750 person years * 10,000 = 105.3 cases/ 10,000 person years

Rate unexposed = 90/24,550 person years * 10,000 = 36.7 cases/ 10,000 person years

c) Rate ratio = 105.3 cases/ 10,000 person years / 36.7 cases/ 10,000 person years = 2.87

ERIC Notebook

Second Edition

Selection Bias

Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Selection bias is a distortion in a measure of association (such as a risk ratio) due to a sample selection that does not accurately reflect the target population. Selection bias can occur when investigators use improper procedures for selecting a sample population, but it can also occur as a result of factors that influence continued participation of subjects in a study. In either case, the final study population is not representative of the target population – the overall population for which the measure of effect is being calculated and from which study members are selected.

Selection bias occurs when the association between exposure and health outcome is different for those who complete a study compared with those who are in the target population.

Example

In a case-control study of smoking and chronic lung disease, the association of exposure with disease will tend to be weaker if controls are selected from a hospital population (because smoking causes many diseases resulting in hospitalization) than if controls are selected from the community.

In this example, hospital controls do not represent the prevalence of exposure (smoking) in the community from which cases of chronic lung disease arise. The exposure-disease association has been distorted by selection of hospital controls.

Sources of selection bias.

Selective survival and losses to follow-up

After enrollment of subjects and collection of baseline data there is usually some loss to follow-up, i.e. when individuals leave the study before the end of follow-up. This biases the study when the association between a risk factor and a health outcome differs in dropouts compared with study participants.

Volunteer and non-response bias

Individuals who volunteer for a study may possess different characteristics than the average individual in the target population. Individuals who do not respond to requests to be studied generally have different baseline characteristics than responders. Bias will be introduced if the association between exposure and a health

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

outcome differs between study volunteers and non-responders.

Hospital patient bias (Berkson's Bias)

Berkson's bias may occur when hospital controls are used in a case-control study. If the controls are hospitalized due to an exposure that is also related to the health outcome under study, then the measure of effect may be weakened, i.e. biased towards the null hypothesis of no association.

Healthy worker effect

Generally, working individuals are healthier than individuals who are not working. Therefore, in occupational exposure studies, where cases (or exposed subjects) are workers, controls (or unexposed subjects) should also be workers, otherwise the association between exposure and the health outcome will tend to be biased towards the null.

Selection Bias

Selection bias will occur as a result of the procedure used to select study participants when the selection probabilities of exposed and unexposed cases and controls from the target population are differential and not proportional. This can occur when exposure status influences selection.

Selection bias will occur in cohort studies if the rates of participation or the rates of loss to follow-up differ by both exposure and health outcome status. Although we seldom can know the exposure and health outcome status of non-respondents or persons lost to follow-up, it is sometimes possible to obtain these data from an external source.

Terminology

Bias: a systematic error in a study that leads to a distortion of the results. (*Target population:* the overall population for which the measure of effect is being calculated, and from which study members are selected.

Loss to follow-up: when individuals leave the study before the end of follow-up.

Medical Epidemiology, Greenberg RS, 1993).

Practice Questions

Answers are at the end of this notebook

1) Researchers are planning to conduct a case-control study of the association between an occupational exposure and a health outcome. The researchers plan to study exposed workers from one factory and compare them with unexposed retirees who have never worked in a factory. A reviewer of the research proposal is worried about selection bias and in particular about the possibility of the healthy worker effect. Which of the following best represents the reviewer's concern?

a) Retirees should not be compared to factory workers because factory workers are under more stress than retirees

b) Retirees should not be compared to factory workers because factory workers' incomes differ from those of retirees

c) Retirees should not be compared to factory workers because factory workers are likely to need to maintain a certain level of health in order to work in a factory while retirees would not necessarily be as healthy

d) Retirees should not be compared to factory workers because factory workers likely live in a different city than the retirees

2) Researchers conducted a prospective cohort study of the association between air pollution exposure and asthma. Some study participants were lost to follow-up (dropped out of the study) over time. The researchers were able to obtain data on the exposure and the health outcome for participants who remained in the study as well as for participants who dropped out of the study. The researchers discovered that the rate of loss to follow-up did not differ when comparing exposed and unexposed groups. The researchers also found that the rate of loss to follow-up did not differ when comparing people who developed asthma and people who did not develop

asthma. Based on this information, which one of the following statements is most likely to be true?

a) Selection bias likely occurred in this study because both exposure groups experienced loss to follow-up

b) Selection bias likely did not occur in this study because exposure status and health outcome status did not influence whether or not people dropped out of the study

c) Selection bias likely occurred in this study because both of the outcome groups (people with asthma and people without asthma) experienced loss to follow-up

d) Selection bias likely did not occur in this study because people cannot choose if they are exposed to air pollution or not exposed to air pollution

References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Cener at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. Answer choice c is correct. The healthy worker effect is a type of selection bias that may occur in occupational exposure studies when the exposed cases are workers but the non-exposed study participants (controls) are not workers. In general, working individuals are healthier than non-working individuals. Health problems may actually be a reason for not working. In addition, retirees are typically older than the working population and may have more age-related health problems.
2. Answer choice b is correct. Selection bias likely did not occur in this study because exposure status did not influence whether or not people dropped out of the study. Furthermore, the health outcome status did not influence whether or not people dropped out of the study. Remember that selection bias may occur in a cohort study if the rate of participation or the rate of loss to follow-up differ by both exposure and health outcome status. Selection bias is not affected by if the exposure is an avoidable exposure or a non-avoidable exposure.