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?

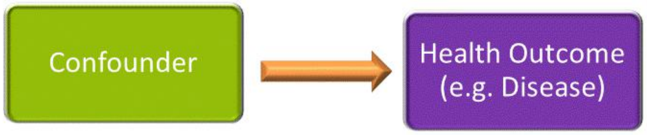
Assessing confounding

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- = 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

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

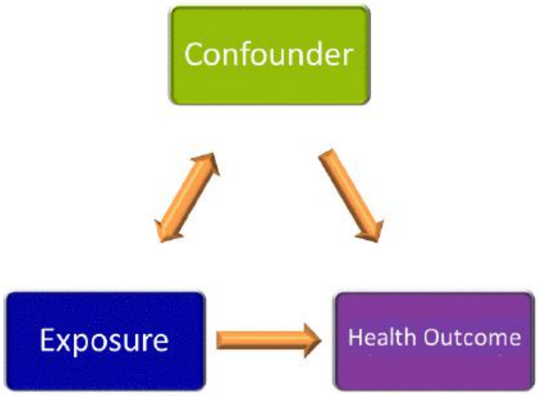
where the odds that the cases have the confounder = the number of cases with the confounder (C+) divided by the number of cases without the confounder (C-) and where population F is comprised of all individuals who are not exposed (E-).

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}=

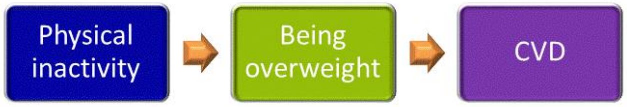
$$\frac{\text{Odds of controls having the confounder (C+) among Population A}}{\text{Odds of controls having the confounder (C+) among Population B}}$$

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.